DOCUMENT RESUME

ED 372 951 SE 054 513

AUTHOR Jorgensen, Margaret

TITLE Assessing Habits of Mind. Performance-Based Assessment in Science and Mathematics.
INSTITUTION ERIC Clearinghouse for Science, Mathematics, and Environmental Education, Columbus, Ohio.
SPONS AGENCY Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE May 94
CONTRACT RI88062006
NOTE 109p.
AVAILABLE FROM ERIC/CSMEE, The Ohio State University, 1929 Kenny Road, Columbus, OH 43210-1080.
PUB TYPE Reports - Research/Technical (143) -- Information Analyses - ERIC Clearinghouse Products (071)

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS *Competency Based Education; Elementary Secondary Education; *Mathematics Instruction; *Portfolios (Background Materials); *Science Instruction; *Student Evaluation
IDENTIFIERS *Performance Based Evaluation

ABSTRACT

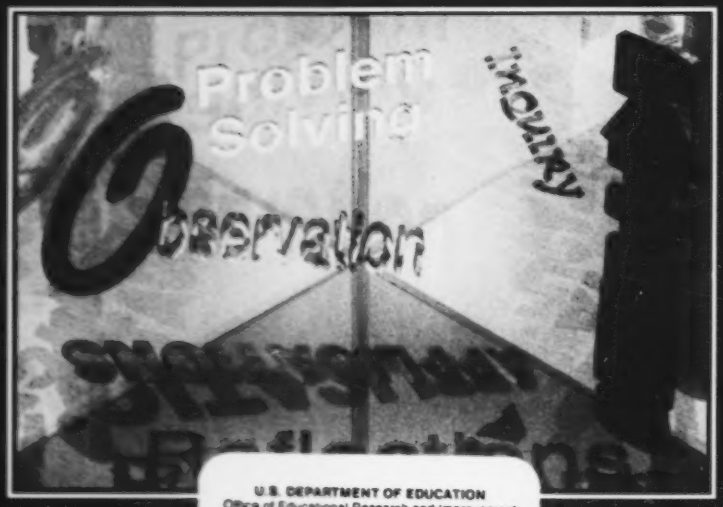
To improve education, educators and policymakers must acknowledge the fact that often what happens in the classroom is governed by achievement tests, college entrance examinations, and other standardized tests. It is for this reason that many have begun to look at methods of assessment when considering the improvement of education. This document presents a discussion of performance based assessment. This method of assessment requires that students complete, demonstrate, or perform the actual behavior of interest. It is noted that the textbook is not to be used as a guide in creating high-stakes assessments for use in promotion/retention, program evaluation, or teacher evaluation; instead it is to be used as a tool for teachers to improve their classroom instruction. The book contains the following chapters: (1) "How Will This Book Help?"; (2) "What Is Performance-Based Assessment?"; (3) "Why Use Performance-Based Assessment in the Classroom?"; (4) "How Can Performance-Based Assessment Really Work?"; (5) "How Can Scoring Guides (Rubrics) Communicate Complex Information?"; (6) "How Can Teachers Be Informed Consumers?"; and (7) "What Are the Critical Questions About Performance-Based Assessment?" Also included is a brief guide to ERIC. (Contains 54 references.) (ZWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Assessing Habits of Mind

Performance-Based Assessment in Science and Mathematics

ED 372 951



U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

By Margaret Jorgensen, Ph.D.

Eric Clearinghouse for Science, Mathematics, and Environmental Education

SE054513

BEST COPY AVAILABLE

Assessing Habits of Mind:
Performance-Based Assessment in Science and Mathematics

Assessing Habits of Mind:
Performance-Based Assessment in Science and Mathematics

Margaret Jorgensen, Ph.D.

May 1994

Produced by the

ERIC

Clearinghouse for Science, Mathematics, and Environmental Education
The Ohio State University
1929 Kenny Road
Columbus, OH 43210-1080

Cite as:

Jorgensen, M. (1994). *Assessing habits of mind: Performance-based assessment in science and mathematics.* Columbus, OH: ERIC Clearinghouse for Science, Mathematics, and Environmental Education.

Document development:

David L. Haury, *ERIC/CSMEE Executive Editor*
Sigrid Wagner, *Publications Coordinator and Copy Editor*
J. Eric Bush and Christine M. Janssen, *Page Layout and Cover Design*

Accession Number: SE054513

This document and related publications are available from ERIC/CSMEE Publications, The Ohio State University, 1929 Kenny Road, Columbus, OH 43210-1080. Information on publications and services will be provided upon request.

ERIC/CSMEE invites individuals to submit proposals for monographs and bibliographies relating to issues in science, mathematics, and environmental education. Proposals must include:

- A succinct manuscript proposal of not more than five pages.
- An outline of chapters and major sections.
- A 75-word abstract for use by reviewers for initial screening and rating of proposals.
- A rationale for development of the document, including identification of target audience and the needs served.
- A vita and a writing sample.

This publication was funded by the Office of Educational Research and Improvement, U. S. Department of Education under contract no. RI-88062006. Opinions expressed in this publication do not necessarily reflect the positions or policies of OERI or the Department of Education.

ERIC and ERIC/CSMEE

The *Educational Resources Information Center* (ERIC) is a national information system operated by the Office of Educational Research and Improvement in the U. S. Department of Education. ERIC serves the educational community by collecting and disseminating research findings and other information that can be used to improve educational practice. General information about the ERIC system can be obtained from ACCESS ERIC, 1-800-LET-ERIC.

The ERIC *Clearinghouse for Science, Mathematics, and Environmental Education* (ERIC/CSMEE) is one component in the ERIC system and has resided at The Ohio State University since 1966, the year the ERIC system was established. This and the other 15 ERIC clearinghouses process research reports, journal articles, and related documents for announcement in ERIC's index and abstract bulletins.

Reports and other documents not published in journals are announced in *Resources in Education* (RIE), available in many libraries and by subscription from the Superintendent of Documents, U. S. Government Printing Office, Washington, DC 20402. Most documents listed in RIE can be purchased through the ERIC Document Reproduction Service, 1-800-443-ERIC.

Journal articles are announced in *Current Index to Journals in Education* (CIJE). CIJE is also available in many libraries, and can be purchased from Oryx Press, 4041 North Central Avenue, Suite 700, Phoenix, AZ 85012-3399 (1-800-279-ORYX).

The entire ERIC database, including both RIE and CIJE, can be searched electronically online or on CD-ROM.

Online Vendors:	BRS Information Technologies, 1-800-289-4277 DIALOG Information Services, 1-800-334-2564 OCLC (Online Computer Library Center), 1-800-848-5800
CD-ROM Vendors:	DIALOG Information Services, 1-800-334-2564 Silver Platter Information, Inc., 1-800-343-0064

Researchers, practitioners, and scholars in education are invited to submit relevant documents to the ERIC system for possible inclusion in the database. If the ERIC selection criteria are met, the documents will be added to the database and announced in RIE. To submit, send two legible copies of each document and a completed Reproduction Release Form (available from the ERIC Processing and Reference Facility, 301-258-5500, or any ERIC Clearinghouse) to:

ERIC Processing and Reference Facility
Acquisitions Department
1301 Piccard Dr., Suite 300
Rockville, MD 20850-4305

ERIC/CSMEE National Advisory Board

Eddie Anderson, Chief, Elementary and Secondary Programs Branch of the Education Division, National Aeronautics and Space Administration

Linda Ruiz Davenport, Center for the Development of Teaching, Education Development Center

James D. Gates, Executive Director, National Council of Teachers of Mathematics

Louis A. Iozzi, Director, Center for Environmental Education, Cook College, Rutgers University

J. David Lockard, Director, International Clearinghouse for the Advancement of Science Teaching, University of Maryland

Phyllis Marcuccio, Assistant Executive Director for Publications, National Science Teachers Association

Kathy McGlauflin, Vice President for Education, American Forest Foundation

Kate Nevins, Vice President for Member Services, Online Computer Library Center

Michele Perrault, Director of Teacher Education Programs, Sierra Club (National President, 1993-94)

Thomas A. Romberg, Director, National Center for Research in Mathematical Sciences Education, University of Wisconsin-Madison

Lyn Taylor, School of Education, University of Colorado-Denver

Table of Contents

Acknowledgments ix

Chapter 1
How Will This Book Help? 1

Chapter 2
What Is Performance-Based Assessment? 9

Chapter 3
Why Use Performance-Based Assessment in the Classroom? 21

Chapter 4
How Can Performance-Based Assessment Really Work? 35

Chapter 5
How Can Scoring Guides (Rubrics) Communicate Complex Information? 47

Chapter 6
How Can Teachers Be Informed Consumers? 59

Chapter 7
What Are the Critical Questions About Performance-Based Assessment? 65

References 73

Appendix A 77

Appendix B 79

Appendix C 81

Appendix D 88

A Brief Guide to ERIC 91

Acknowledgments

Writing a practical "how to" book for classroom teachers in the area of performance-based assessment is the natural outgrowth of my work at Educational Testing Service (ETS). ETS has provided me with the opportunity to experiment with ideas and concepts, to explore new techniques, and to search for a structure to guide assessment development in this new and exciting area of measurement. Among the many supporters across ETS, two individuals have been particularly helpful to me in this endeavor. I want to take this opportunity to formally thank them.

Dr. Roy Hardy, Director of the Southern Field Office of ETS, started me on the "road to performance-based assessment" in 1989 when he included me in a pivotal project working with Georgia teachers in the development of innovative assessments to inform instruction. It was this project which sparked my interest in staff development in the area of performance-based assessment. Ms. Martha-Anne (Marty) McDevitt, Associate Examiner in the Southern Field Office, has been my sounding board and co-presenter in many of my staff development programs in performance-based assessment. Their cooperation, challenges, and fruitful questions have helped me clarify many important issues discussed in this text.

Many other ETS colleagues have given "food for thought" as I have continued to train teachers in the development of performance-based assessment. My sincere thanks go to Ms. Roberta Camp, Dr. Edward Chittenden, Dr. Nancy Katims, Dr. Stephen Koffler, Dr. Terry Salinger, and Mr. J.T. Stewart. A special thanks to Ms. Drucilla Jackson for her help in preparing the manuscript and to Ms. Katherine Goodman for her editorial assistance.

Finally, I am grateful to Joan Boykoff Baron, Douglas A. Rindone, and the Connecticut Department of Education for their willingness to allow the use of "The Soda Task" in this text. "The Soda Task" is an example of high quality performance-based assessment developed by the Connecticut Department of Education, as part of a National Science Foundation funded research project (SPA 8954692) on performance assessment in mathematics and science. A book entitled *Connecticut's Performance Assessment in Mathematics and Science: Telling the Whole Story* (J.B. Baron, Ed.) that describes this project, including various examples of performance tasks, scoring guides, and students' work, is in progress.

Chapter 1
How Will This Book Help?

> *One of the most common criticisms of education today is that schools spend too much time stuffing kids with information and too little time teaching them how to think. ... Many teachers don't like this approach but say an important part of their job is helping children learn the "stuff" they need to know to pass achievement tests. ... It's the standardized test results that often determine whether a student will advance a grade, graduate or go to college. So until the tests themselves change, teachers say, school curriculum must follow their lead, at least in part.* (McCartin, 1992, p. E1)

Achievement tests play a powerful and prescriptive role in influencing education. But change is afoot. In fact, the buzzword in American society today is *change*—change in politics, change in economics, and change in education. In education, change is being advocated not only in curriculum and instruction but also in assessment.

A sense of urgency grips this country with the recognition of clear and consistent evidence that American education does not produce *thinking* individuals. Data from the National Assessment of Educational Progress (NAEP) testify to the fact that the achievement record in science, mathematics, reading, and writing declined during the seventies and then rebounded to the 1970 baseline during the eighties (Mullis, Dossey, Foertsch, Jones, & Gentile, 1991). In short, after years of concern about the state of education in this country, achievement levels have not risen above the relatively mediocre levels recorded twenty years ago.

The challenge is clear. Students need to learn to be thinkers and they must be able to demonstrate what they have learned in meaningful ways. Gregory Anrig, past president of Educational Testing Service, comments on the consequences of this critical situation:

> Children are not learning enough for the world that awaits them. ... The world is not waiting for us. In a recent International Assessment of Educational Progress, the mathematics and science achievement of 9- and 13-year-old students in 15 countries was compared. Except for the science achievement of 9-year-olds, U.S. children came out close to the bottom. Twenty years and more from now, today's 13-year-olds will be sitting across the economic table trying to negotiate with contemporaries from Asia and Europe whose knowledge will outgun them if we don't get our educational act together quickly. (Anrig, 1992, p. 1)

While the challenge may be clear, the solution is not. It is certain, however, that assessment will play a visible and powerful role in changing the shape of American education to create a *generation who thinks*.

In September 1989, President Bush and the nation's governors met in Charlottesville, Virginia, for an Education Summit. The report produced at this meeting is often referred to as the Jeffersonian Compact. This document provided the political force that put education reform into motion. At this meeting, President Bush and the governors reached agreement on six National Education Goals. President Bush formalized this education plan in *America 2000*, in which the goals adopted at the Education Summit were highlighted:

By the Year 2000:

1. All children in America will start school ready to learn.
2. The high school graduation rate will increase to at least 90 percent.
3. American students will leave grades four, eight, and twelve having demonstrated competency in challenging subject matter, including English, mathematics, science, history, and geography; and every school in America will ensure that all students learn to use their minds well, so they may be prepared for responsible citizenship, further learning, and productive employment in our modern economy.
4. U.S. students will be the first in the world in science and mathematics achievement.
5. Every adult American will be literate and will possess the knowledge and skills necessary to compete in a global economy and exercise the rights and responsibilities of citizenship.
6. Every school in America will be free of drugs and violence and will offer a disciplined environment conducive to learning. (U.S. Department of Education, 1991, p. 3)

These goals have been reiterated and encouraged by President Clinton who, as Governor of Arkansas, was a leader in their original formulation and adoption. In *Goals 2000: Educate America Act* (1993), President Clinton calls for fundamental reform in schools and school systems throughout the country—for challenging curriculum standards, better assessments, and more opportunities for students to meet high standards.

Although the "projection of education as a vital national concern is probably the most important, substantive and symbolic consequence of *America 2000*" (Chira, 1991, p. 1), implementing effective reform will be difficult. As Chira goes on to say:

> Making America an educational as well as a military superpower will mean confronting several crises: the glaring failure of the worst students, the tolerance of mediocrity and a national heritage of anti-intellectualism. (p. 1)

However, the strength of bipartisan support for change forged at the Education Summit has elevated the problems facing education in American to a shared societal platform. In fact, the problems of education are at the root of many other societal problems and, as such, warrant a discussion and action platform not tied to parties in power.

Goal 4, which calls for U.S. students to be first in the world in science and mathematics achievement by the year 2000, has mobilized and legitimized reform in these content areas. Science and mathematics educators have taken the lead, both in describing the learning behaviors desired in American students and in defining standards for curriculum content. *Science for All Americans* calls for students with scientific "habits of mind" (American Association for the Advancement of Science [AAAS], 1989, p. 133). *Everybody Counts* calls for schools to produce workers who are "mentally fit" (National Research Council, 1989, p. 2).

Calls for students adept at higher-order thinking, critical thinking, and problem solving are not restricted to discussions in scholarly publications or education journals read primarily by school administrators and classroom teachers. Instead, they are on the lips of politicians, representatives of business and industry, and policymakers.

As education reform begins to take hold across the country, the role of assessment in effecting change is becoming more and more clear. Classroom instruction will not focus on higher-order thinking skills as long as traditional multiple-choice tests measure primarily low-level recall and recognition skills. Thus, an important aspect of reform is developing innovative methods of assessment.

It has long been acknowledged that the content and structure of multiple-choice tests influence what happens in the classroom. For example, if a state mandates the use of standardized multiple-choice tests in specific content areas at certain grade levels, there is a high probability that those content areas will be emphasized in instruction at those grade levels. An excellent example of how this reality has benefited certain areas of instruction is in language arts, where writing has become a part of the language arts curriculum largely because students' writing skills are directly evaluated by many statewide testing programs. On the other hand, the majority of testing programs do not test science in the elementary grades, so this content area is often neglected in the daily process of instruction (Madaus et al., 1992).

But times are changing. "Improving student learning in mathematics and science is a high priority for our elementary and secondary schools" (Blank & Dalkilic, 1992, p. 1). *A Nation at Risk: The Imperative for Educational Reform* (National Commission on Excellence in Education, 1983) deplored the "rising tide of mediocrity" in the American education system, identifying, in particular, weaknesses in science and mathematics. Subsequent to this report, "virtually every state approved policy initiatives aimed at improving the quality of education" (Blank & Dalkilic, p. 2). The impetus for reform in science and mathematics has been further strengthened by the publication of *Science for All Americans*, the *Curriculum and Evaluation Standards for School Mathematics* (National Council of

Teachers of Mathematics [NCTM], 1989), and the *Professional Standards for Teaching Mathematics* (NCTM, 1991).

In 1991-92, the Council of Chief State School Officers (hereafter referred to as Chiefs) initiated a study of state policies in science and mathematics. They surveyed all state supervisors of science and mathematics in December 1991, seeking information on policies relative to graduation requirements, student assessment programs, and teacher certification. Of particular interest here are the data on curriculum frameworks and assessment.

In mathematics, the Chiefs' study (see Blank & Dalkilic) indicates that 41 states have revised or are revising their state curriculum frameworks based on the NCTM *Standards* and 4 more states are initiating the development of such frameworks (see Appendix A). In science, results indicate that state frameworks exist in 30 states, with 15 other states currently developing such frameworks (see Appendix B). However, unlike mathematics, where frameworks are explicitly based on the *Standards,* it is not clear from the survey the extent to which the science frameworks represent reform in the spirit of *Science for All Americans.*

When questioned regarding the relationship between the structure of the state frameworks and the structure of statewide assessments, the response in mathematics was that in 22 states the assessment program has a *direct* relationship to the curriculum framework, meaning that the "state curriculum framework or guide defines the content topics and skills to be assessed in mathematics" (Blank & Dalkilic, p. 7). Ten states report an indirect relationship, and 5 states have a policy mandating learning outcomes, an important philosophical position that supports innovative assessment (Appendix A). In science, state assessments have a direct relationship to curriculum frameworks in 16 states, an indirect relationship in 7 states, and 6 states have a policy that mandates learning outcomes (Appendix B).

Data on state-mandated tests in science and mathematics were reported as follows (see Appendix C):

- 27 states require a science achievement test (unchanged since 1989)
- 46 states require a mathematics achievement test (up by 6 states since 1989)
- 5 states require a science competency test (down by 1 state since 1989)
- 21 states require a mathematics competency test (up by 2 states since 1989).

(Blank & Dalkilic, p. 13)

In terms of alternative or innovative assessment practices, the Chiefs' study found that, as of spring 1992, 20 states were designing, piloting, or implementing some form of alternative assessment in mathematics, and 16 states were doing the same in science (see Table 1). Assessment formats span a continuum of constructed responses, from so-called enhanced multiple-choice to extended performance. Also varied is the degree of implementation, with 12 states engaged in every-pupil innovative assessment, 14 states engaged in statewide sampling, and 7 states still in the design phase.

Table 1. States With Alternative Assessments in Mathematics and Science

State	Subject/Grade	Type of Assessment	Status
ALABAMA	Algebra 1	Performance	5
ARIZONA	Math 3,8,12	Performance	3,4
	Science 3,8,12	Performance	1,2
CALIFORNIA	Math 4,8,10	EMC, Open-Ended	4
	Science 5,8,10	EMC, Open-Ended	4
COLORADO	Math 4,7,10	Performance, EMC, EPER, Projects	3
CONNECTICUT	Math 10,11	Performance, Open-Ended	1,2,3
	Science 10,11	Performance, Open-Ended	1,2,3
DIST. OF COLUMBIA	Math 7-12	Performance	3
	Science 7-12	Performance	3
FLORIDA	Planning	—	—
GEORGIA	Planning	—	—
HAWAII	Math 4,8	Performance, EMC	4
	Science 3,6,8,12	Performance, EMC	3,4
INDIANA	Math 10	Performance	4
KANSAS	Math 4,7,10	Performance, EMC 1,2,3	3,4
		Open-Ended	5
KENTUCKY	Math 4,8,12	Performance, EMC, Open-Ended	1,2,3
	Science 4, 8, 11	Performance, EMC, Open-Ended	1,2,3
MAINE	Math 4,8,11	EPER, EMC	4,5
	Science 4,8,11	EPER, EMC	4
MARYLAND	Math 3,5,8	Performance	5
	Science 3,5,8	Performance	5
MASSACHUSETTS	Math 4,8,12	Open-Ended	5
	Science 4,8,12	Open-Ended	5
MINNESOTA	Math 5,8,11	Open-Ended, EMC	3,4
	Science 6,9,11	Performance, EMC	4
MISSOURI	Science 7	Performance	3
NEW JERSEY	Math 8,11	Performance, Open-Ended	4,5
NEW YORK	Science 4	Performance	5
NORTH CAROLINA	Math 3-8	Performance	5
	Science 3-8	EMC, Open-Ended	1,2,3
OREGON	Math 8	Performance	5
PENNSYLVANIA	Planning	—	—
TEXAS	Science 9	EPER	3
VERMONT	Math 4,8	Portfolio	4
VIRGINIA	Math	Portfolio, Performance, Projects	1
WEST VIRGINIA	Math 1-6	Performance, EMC	5
	Science 1-6	Performance, EMC	5
TOTAL	Math = 20 states Science = 16 states		

Types of Assessment: Portfolio, Performance, Enhanced Multiple Choice (EMC), Open-Ended, Extended Performance (EPER), Projects. Status: (1) Design, (2) Being written, (3) Being tried out, (4) Being used on statewide sampling basis, (5) Every-pupil basis.

Source: State Departments of Education, Assessment Directors, Fall 1991, and Science and Mathematics Supervisors, Spring 1992.

It is reasonable to interpret the results of the Chiefs' survey in two different ways:

1. Innovative assessment practice in mathematics and science has a strong basis in state policy and clearly indicates the future direction of assessment in these content areas.
2. The policy base for innovative assessment reflects caution. Rather than moving toward full and widespread implementation, policymakers are waiting for results from research to justify the shift from an assessment paradigm defined by choice selection to one governed by production.

From the perspective of cautious commitment to change, it is useful to ask the question, "What is good assessment?" That is, what is assessment in the service of education reform? Herman (1992) suggests that:

> Good assessment is built on current theories of learning and cognition and [is] grounded in views of what skills and capacities students will need for future success. To many, good assessment is also defined by what it is not: standard, traditional multiple-choice items. (p. 75)

Herman goes on to point out what those current theories of learning are:

> According to cognitive researchers, meaningful learning is reflective, constructive, and self-regulated. To *know* something is not just to have received information but to have interpreted it and related it to other knowledge one already has. (p. 75)

It follows that *good* innovative assessment must also be reflective, constructive, and self-regulated. Clearly, these adjectives do not describe typical standardized, norm-referenced or criterion-referenced, multiple-choice tests.

Just as education reform is multidimensional, so is innovative assessment. Chittenden (1990) has said that the labels used today to describe innovative assessment—*authentic assessment*, *alternative assessment*, *portfolio assessment*, *curriculum-embedded assessment*, and *targeted assessment*, among others—are placeholders—that authentic assessment, for example, is more a wish than a technical term. Performance-based assessment also suggests a wish, a belief, a philosophy of learning, and a commitment to change in education that may bring discomfort and require struggle for teachers, students, and, most certainly, for test developers.

What these labels mean in terms of innovation and how they differentially describe innovations are not completely clear. What is clear, however, is that these innovative assessments are *not* traditional multiple-choice tests. And part of education reform is indeed a rejection of the exclusive use of multiple-choice tests in favor of a variety of assessments that will support an educational environment

in which students think and develop and use scientific habits of mind. The goal is to nurture students who are mentally fit for the opportunities of the future. To help teachers move beyond rhetoric to a substantive understanding of innovative assessment is the purpose of this book.

Policymakers have provided the mandate for change in assessment practice. Educators have the motivation. But, before policy and motivation can shift the assessment paradigm in significant ways, it is essential that there be widespread understanding and acceptance of the new way of thinking about "tests."

This chapter has set the stage for using performance-based assessment as a tool in support of a thinking educational environment. Belief in performance-based assessment does not, however, provide the necessary strategies for identifying, developing, and using performance-based assessment. In the chapters that follow, these strategies are presented. The reader will be guided through the design, development, scoring, and interpretation of performance-based assessments.

Throughout the chapters that follow, the reader should keep in mind that this book is intended primarily as a tool for teachers to use to inform their classroom instruction. This book is *not* intended as a guide in creating high-stakes assessments for use in promotion/retention, program evaluation, or teacher evaluation. Performance-based assessments used in high-stakes situations require a more rigorous and systematized approach to development and implementation than has been attempted here.

Chapter 2
What Is Performance-Based Assessment?

This chapter presents the rationale for performance-based assessment, establishes the relationship between performance-based assessment and other kinds of innovative assessment, and discusses implications of the paradigm shift in assessment.

The United States has a history of efforts at school reform. In the late 1950s and early 1960s, Sputnik and the race to the moon stimulated reform in mathematics and science education. In the late 1960s and early 1970s, education reform was stimulated by the Great Society and a concern for equality of opportunity more than for international competitiveness. In the late 1980s and early 1990s, education reform has again been stimulated by competition and economics. But there is a vast difference in the way reform is being positioned in the 1990s, and that difference has to do with the way in which education is being evaluated.

Traditionally, education has been viewed as a system of inputs and outputs. The inputs are defined in terms of human and capital resources, instructional programs, physical facilities, and expenditures. As such, the reform movements of the past have focused on input variables, such as per-pupil expenditures, class size, number of books in the library, teacher tenure and credentials, school time (hours and days), facilities, available technology, instructional materials, and student and teacher absenteeism. The products of the educational system have traditionally been defined in terms of output variables, such as attendance, retention rate, graduation rate, matriculation into higher education, and test scores.

With this understanding of the process of education as one defined by the inputs and evaluated by the outputs, reform in education has had these same foci. As a result, reform has meant more dollars, more books, new facilities, computers in the classrooms, and so forth. And, the evaluation of reform has examined output variables. In short, "to improve education meant to try harder, to engage in more activity, to magnify one's plans, to give people more services, and to become more efficient in delivering them" (Finn, 1990, p. 584).

When citing the need to look at problems from new perspectives as advocated by Thomas Kuhn in *The Structure of Scientific Revolutions* (1970) or by Joel Barker in *Future Edge* (1992), the traditional paradigm that defines education as a system of inputs and outputs seems clearly inadequate, inefficient, and unproductive. The new paradigm for education is one that focuses on the development of intellectually competent people as the products of the educational system. No longer are output variables like attendance, for example, the measure of success. The only meaningful products of the United States educational system are individuals who "use their minds well" (U.S. Department of Education, 1991).

The new paradigm acknowledges that there is not likely to be one system, one process, which works to produce students who can think, who have scientific habits of mind, and who are mentally fit. In fact, the more that is known about instruction, learning behavior, learning styles, and intelligence, the less reasonable it is to impose a single process for learning on students. Likewise, it is presumptuous to impose a process for assessment on students that does not recognize individuality. It is this somewhat radical paradigm shift that opens the door to innovative assessment in the service of education.

With this paradigm shift, the following changes are likely to occur:

Traditional Paradigm	Reform Paradigm
Teacher-Led Instruction	Student-Led Instruction
Segmented School Schedule	Flexible Schedule
Rigid Scope and Sequence Curriculum	Optional Modular Curriculum
Separation of Content Domains	Thematic Instruction
Traditional Assessment	Performance-Based Assessment

The shift from the input-output model of education to the intellectual competence model has primed the pump for changes in methods of assessment.

The Role Of Paradigms

Before moving ahead, it is helpful to understand the full meaning of paradigms and the tremendous forces required for a shift in paradigms. Barker (1992, p. 32) defines a paradigm in the following manner:

> **Paradigm**—A set of rules and regulations that does two things:
> (1) it establishes or defines boundaries
> (2) it tells you how to behave inside the boundaries to be successful.

It is important to understand the tremendous resistance to innovation that faces the education reform movement in order to understand the courage and creativity that will be required, not only to move beyond multiple-choice tests to

more informative assessments, but also to explore and use these innovative measures, should the information gleaned be either unflattering to educators or inappropriate in specific situations. In short, there is risk in change. And with the paradigm shift from traditional multiple-choice tests to other forms of assessment, all the risks inherent in discovery and adventure are present.

The risk taker—the paradigm pioneer—must look for rules and regulations that will eventually redefine the boundaries of education. The paradigm pioneer must think creatively about solutions rather than obstacles. The paradigm pioneer must look to other environments in which the quality of the outputs is more important than standardization of the process. Perhaps foremost, the paradigm pioneer must take chances, must try things that *have never been done before*, must take strategies traditionally used in one context and broaden their application. This creative and risky business of change is both challenging and threatening.

There is some comfort, however, in the realization that the vision of a generation of citizens who truly think, who use their minds well, and who are mentally fit is the only meaningful product of the American educational system. Thus, the spirit, the philosophy underpinning this revolution—this paradigm shift—is well worth the risk and effort inherent in change.

Evidence of the Paradigm Shift

What is the paradigm shift in assessment? What are innovative assessment techniques and strategies? How will these innovations in assessment contribute to the paradigm shift in the teaching-learning interaction in schools across this country?

- *The most visible evidence of the shift to a new assessment paradigm is the plethora of labels describing this new assessment perspective.*

The search for innovative assessments has led to "confusion in packaging." This confusion stems as much from enthusiasm and diversity as from methodological differences. In searching for assessments to support the vision of a thinking generation of students, the alternatives to traditional multiple-choice tests may be called alternative assessment, authentic assessment, curriculum-embedded assessment, portfolio assessment, or targeted assessment. Precisely what is meant by these labels is unclear in many instances. There are probably as many definitions for these and other labels as there are measurement experts and educators addressing innovative assessment.

Because of (a) the widespread interest in innovative assessment; (b) the relative scarcity of expertise, strategies, procedures, and guidelines; and (c) the small but growing collection of assessments to model, the work emerging often reflects more the individual preferences of the authors/developers than consensus within the measurement community. The coordination and scrutiny that have contributed to confidence in the design and use of standardized multiple-choice tests have not yet come into play in the area of innovative assessment. In order to

move forward effectively, the individuals working on innovative assessment must develop a common vocabulary.

The theme of this book is performance-based assessment. What, then, does it mean for an assessment to be performance-based? One definition is:

> Performance-based assessment requires that the student complete, demonstrate, or perform the actual behavior of interest. There is a minimal degree of inference involved.

For example, if the behavior of interest is writing, the student actually writes. The student does not complete multiple-choice questions about sentences and paragraphs, about punctuation and mechanics, or about syntax and tone. If the behavior of interest is scientific investigation, the student conducts a scientific investigation. The student does not answer multiple-choice questions about steps in the scientific method, definitions of terms, or the setup for a prescribed experiment. If the behavior of interest is mathematical reasoning, the student engages in mathematical reasoning. The student does not answer multiple-choice questions about algorithms, strategies, or computations.

If students were given multiple-choice questions about writing, scientific inquiry, or mathematical reasoning, the assessment would require some degree of inference about the transfer from discrete, *chunkable* skills to more generalizable behaviors that reflect real life and the world of work. The examiner, test developer, or evaluator must draw an inference from behavior on isolated chunks of ideas to performance on a larger, more complex whole.

- *The paradigm shift in assessment reflects corresponding shifts in philosophy and learning theory.*

Most notable in the paradigm shift in the philosophy and theory of learning is the emphasis on performance of complex, holistic tasks rather than "snippets" of performance in segments, elements, or chunks of complex tasks. Inherent in the assessment paradigm shift is a belief that complex learning behaviors cannot be decomposed into independent bits of knowledge and skills that can then be tested and the results combined to reflect the larger complex behavior. Resnick and Resnick (1989) suggest that this belief in the indecomposability of complex competencies is very much at the heart of the innovative assessment movement.

Shepard (1991, p. 2) writes about the same issue but from the perspective of measurement theory as a reflection of learning theory. She writes that the conceptions of teaching and learning invoked by measurement specialists when they structure multiple-choice tests may run counter to what is currently known about learning. In short, if traditional multiple-choice tests derive from behaviorist learning theory, which requires sequential mastery of constituent skills and behaviors, then these tests are inappropriate for evaluating learning that is not sequential, hierarchical, or decomposable.

◆ *The paradigm shift in assessment reflects a changing emphasis in curriculum content.*

There appears to be a clamor for *hard content* emerging as a goal.

> Hard content means not just the facts and skills of academic work, but understanding concepts and the interrelationships that give meaning and utility to the facts and skills.... The emphasis is on students learning to produce knowledge, rather than simply reproduce knowledge.
>
> (Porter, Archbald, & Tyree, 1991, p. 11)

The focus is not just on hard content for the college-bound student. Instead, the demographics of the work force and the changing nature of the world of work necessitate that all students experience hard content and that they rise to the challenge of being active and enthusiastic learners throughout their lifetimes.

Curriculum standards that meet the emerging definition of hard content are beginning to appear in the literature, with the National Council of Teachers of Mathematics (1989) in the lead, and other professional organizations not far behind. What remains is for educators to adopt, amend, and implement these content standards, and then to develop assessments and articulate performance standards for use in the assessments. These are not easy tasks, for they involve answering questions like "What does it mean to 'use the mind well,' to reason, to communicate effectively?"

Tests used to support and encourage the learning or solution of complex tasks must themselves be complex tasks. This philosophical position is supported by cognitive researchers (e.g., Resnick & Resnick, 1989). Learning is no longer viewed as a process by which students master hierarchically ordered skills that culminate in a complex outcome. Learning is viewed as the progressive refinement of the combination of complex skills applied to rich content.

Another way to describe the paradigm shift in assessment is to identify the shifting elements shown at the top of the next page.

These points of contrast explain quite clearly the fervor and energy directed towards performance-based assessment; traditional assessment is, quite simply, contrary to the goals of education reform.

◆ *The paradigm shift is toward complex assessment.*

Complex assessment tends to have the following characteristics:

- It uses learning tasks as source and resource.
- It involves an extended time frame.
- It can involve group activities.
- It uses process as an evaluation criterion.
- It requires human raters to make decisions about the performance.

(Baker, 1990)

Performance-based assessment is, by its very nature, complex assessment.

Traditional Assessment	Performance-Based Assessment
Controlled Time for Administration	Variable Time for Administration
Individual Effort	Individual Effort and/or Collaborative Effort
Controlled by Developer/ Administrator	Controlled by Students
Emphasis on Answer/Output	Emphasis on Process/Product
Content is Focused and Discrete	Content is Broad and Holistic
One Correct Answer	Multiple Correct Answers
Response Mode is Fixed	Response Mode is Selected by the Student
Performance Standards Are Empirically Derived	Performance Standards Are Derived From an Understanding of the Content

Other Kinds of Innovative Assessment

Some of the other kinds of innovative assessment already mentioned are alternative assessment, authentic assessment, and portfolio assessment. Baker describes *alternative assessment* as anything that is not multiple-choice (or other format that requires only a selection from a list of choices) and *authentic assessment* as heavily contextual. Meyer clarifies this with the following definition:

> In an authentic assessment, the student not only completes or demonstrates the desired behavior, but also does it in a real-life context. (Meyer, 1992, p. 40)

Zessoules and Gardner (1991) offer a more elaborate definition when they add that performance criteria may be stated in terms of the student's classroom world or an adult expectation. They also suggest that the significant criteria which document the authenticity of the performance must be clearly identified. This

extension of assessment into the real world is an appealing attribute. At the same time, the more real-world the assessment, the less clear the information from that assessment will be in terms of who or what contributed to what was observed.

Grant Wiggins (1990) defines authentic assessment somewhat differently:

> Assessment is authentic when we directly examine student performance on worthy intellectual tasks.... Authentic assessments require students to be effective performers with acquired knowledge. Authentic assessments present the student with the full array of tasks that mirror the priorities and challenges found in the best instructional activities.... Authentic assessments attend to whether the student can craft polished, thorough and justifiable answers, performance or products.... Authentic assessment achieves validity and reliability by emphasizing and standardizing the appropriate criteria for scoring such products.... Authentic tasks involve 'ill-structured' challenges and roles that help students rehearse for the complex ambiguities of the 'game' of adult and professional life. (p. 20)

In fact, these characteristics will be re-examined in Chapter 4. At this point, however, the distinction between authentic and performance-based assessment that may be most useful is that authentic assessment is a subset of a broad assessment arena that requires performance or demonstration of complex cognitive behaviors. To the extent that these assessment opportunities are set in the real world, they may indeed be authentic as well as performance-based.

Portfolio assessment is another term worthy of discussion. Salinger (1992) defines *portfolio* as follows:

> **Portfolio**—A purposeful collection of student work that exhibits the student's efforts, progress, and achievement in one or more areas. Words used to describe these purposeful collections often include *collection, selection,* and *reflection* to emphasize the interactive process among the teacher, learner, and materials.

The reflection feature is common across various types of innovative assessments. It is also important to realize that the goal of creating an assessment which closely imitates the learning outcome of interest is not new to performance-based assessment, although from the rhetoric, it would seem that notion is totally revolutionary. As long ago as 1951, E. F. Lindquist (the creator of the first optical mark-sense reader and the principal author of the *Iowa Test of Basic Skills*) cautioned that an achievement test developer should always construct items as

similar as possible to the criteria being measured. Thus, it is not that traditional assessment methods do not attempt to match the criteria of interest; it is that reform has redefined the criteria to be beyond the capabilities of the multiple-choice format.

Testing and Assessment

With this background in labels for innovative assessment, it is now appropriate to examine the characteristics or properties that move the activities described above from the instructional arena to the measurement arena, in which assessment properly belongs. As performance-based assessment emerges as a field in its own right, it is important to identify those aspects of traditional measurement theory that remain pertinent to innovative assessment and apply them in this new paradigm. At a minimum, the lessons learned in the field of traditional assessment should be used to inform practice in areas of innovative assessment.

Again, terminology is important. What is the difference between *test* and *assessment*? The common understanding of the word *test* in the measurement community is similar to that in the medical world—a test is a single procedure that provides information which, in turn, provides the basis for decision making (e.g., diagnosis and prescription in the medical arena). A common understanding of the word *assessment* is that of a system of procedures that provides the basis for decision making. So, for example, a test to sample achievement in mathematics would likely look the same for each examinee (i.e., everyone would take the same set of questions or parallel sets of questions with each question having 4 or 5 options from which to choose the correct answer), whereas an assessment would comprise a variety of tests to measure or document the behavior. These tests would presumably reflect different perspectives, modalities, or structures.

In terms of mathematics, a test designed to capture problem-solving skills in the area of fractions might consist of fifty multiple-choice questions. Some of these questions might incorporate situations or scenarios, data to be analyzed, or transformations from one representation to another. These questions might tap higher-order thinking skills if defined as "application" and above (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956). However, options would generally be presented from which to select the correct answer to each item.

Moving from multiple-choice testing to performance-based assessment requires more than abandoning options for answers. It is not a paradigm shift to have 50 questions measuring problem solving in fractions with a line on which the student writes the correct answer.

Consider the following:

Sam wants to buy enough of the same fabric to make matching vests for himself and two of his friends. Each vest requires three-fourths of a yard of fabric. He likes two different fabrics. The plaid fabric costs $ 3.45 per yard. The striped fabric costs $ 2.87 per yard. Select the fabric you would prefer and tell how much it will cost to make the three vests.

Answer each of the following questions:

1. Which fabric would you have Sam choose for the vests?

   _____ fabric

2. How much will the fabric cost for each vest for your choice of fabric?

   $ _____ cost of fabric for 1 vest

3. If Sam adds $ 1.78 for the cost of buttons, thread, and lining for each vest, what is the total cost for each vest?

   $ _____ total cost for each vest

4. How much will all three vests cost Sam?

   $ _____ total cost for all 3 vests

This is an example of a multi-step, *constructed response* test question that might be classified as a performance test question because it requires that the examinee interpret and relate information provided in the question with prior knowledge. However, it is structured in that the examinee (student) is told how to proceed from the presentation of the problem to the solution. Furthermore, there is little flexibility with regard to response. There are two choices for fabric and one correct dollar amount for each fabric choice. Given these constraints and this structure, this test question is more consistent with the traditional assessment paradigm than it is innovative. Although the student does have an opportunity to *produce* rather than *select* a cost per vest, this testlet could easily be replaced with multiple-choice questions without sacrificing significant information.

On the other hand, suppose the question were to read as follows:

Sam wants to buy enough of the same fabric to make matching vests for himself and two of his friends. Each vest requires three-fourths of a yard of fabric. How much will it cost to make these three vests?

Think about this task and use the "Fabric World" advertisement below to select the fabric or fabrics you would use. Be prepared to show what you thought about, what decisions you made, and why. Select an effective way to communicate this information to your teacher and classmates.

Fabric World SALE!

100 % Cotton, lightweight, prints and solids
Regularly $ 4.50 per yard, NOW $ 2.25

Denim, prewashed, heavyweight, latest shades
Regularly $ 8.00 per yard, NOW $ 4.00

Sail Cloth, colorfast, stripes and solids
Regularly $ 6.00 per yard, NOW $ 3.50

Answer:

[More space would be provided]

The lack of constraints or restrictions on how the problem should be solved allows students to approach the problem differently, to integrate knowledge and processes in a way deemed appropriate to each student, and to demonstrate the common big ideas of reasoning, problem solving, communication, and connections. This openness or opportunity for each student to shape the problem and solve it from his or her own perspective and knowledge base is a hallmark of performance-based assessment and is not possible in a traditional testing situation.

Another view of the difference between testing and assessment comes from Bloom, Hastings, and Madaus (1971). They describe assessment as being multidimensional in nature. That is, an assessment would use various tools to measure the same behaviors (outcomes). So, for example, a personality assessment would require multiple measures, only in combination defining the complex phenomenon called *personality*.

In terms of labels, the distinction between test and assessment may be subtle. As a statement of underlying rationale, however, the notion that a tool used to describe learning within the context of the classroom should be relatively

unstructured, unconstrained, and supportive of the individual response preferences of different students is an important characteristic to remember in the construction and selection of performance-based assessments. This idea will be discussed in more detail in Chapter 4.

Implications of the Paradigm Shift

In performance-based assessment, just as in traditional test development, there must be a strategy for systematically observing behavior and describing it with a numerical scale or category system, or there must be some other method of synthesizing and summarizing observed behavior for the purposes of communication and analysis.

With both performance-based assessments and traditional testing situations, it is reasonable to expect debates over the relative usefulness of different measurement strategies and the ever-present quest for that one instrument, that one test, which will do everything for everyone. Just as when norm-referenced tests are reinterpreted so that criterion-referenced information can be made available to users, and when criterion-referenced tests are normed, so too are performance-based assessments developed for use in the classroom likely to be co-opted into high-stakes assessment programs. Or, conversely, high-stakes assessments or their look-alikes may be borrowed and embedded in classroom instruction. Each of these misappropriations jeopardizes the utility of the information obtained because the purpose for which each assessment was developed is not that for which it is being used.

It seems that users of tests have always wanted to minimize testing time but, at the same time, to maximize the information available. What has been learned over and over again since the early days of sophisticated measurement practice, dating from the work of Binet and Simon, is that tests designed for specific purposes can be made to do that work well, that is, with objectivity, reliability, and validity. The more specific and well-defined the purpose, the better honed the test can be. The flip side is that, as the test becomes broader in purpose and less well-defined in its focus, the information becomes harder to interpret because so many different variables play roles in influencing performance.

For example, if a 20-question test is designed to measure only two-column addition without regrouping, one can feel relatively comfortable that the results can be interpreted in terms of skill in that area. If, on the other hand, a 20-question test were designed to measure two-column addition without regrouping, and two-column addition with regrouping, and one- and two-step problem solving, and geometric problem solving, the interpretation of the results would become less straightforward. The more broad-based the dimensions of the test become without its being lengthened, the less likely that direct interpretations about specific capabilities can be made with confidence.

Another useful perspective relevant to this discussion is historical. In *Essentials of Psychological Testing*, Cronbach (1970) describes tests of typical performance. These tests are intended to study an individual when he or she is "acting naturally" (p. 39). Cronbach goes on to suggest that observations of

natural behavior (even captured on videotape) can be made in both standardized and unstandardized or "natural" conditions. His examples focus on children interacting with each other. Cronbach also mentions use of self-report devices to collect information on typical performance. Imbedded in this discussion are issues of scoring reliability, or what Cronbach refers to as the dichotomy of psychometric versus impressionistic testing.

Cronbach's book was published over two decades ago. Have we come full circle? If so, what has been learned from objective, standardized, fill-in-the-bubble testing that can help us avoid some of the concerns related to scoring and interpretation, such as accuracy of reporting, objectivity of scoring, standardization, fair and equitable testing situations, the trade-offs between detailed observation and recording, and the efficiency of testing, scoring, and reporting results for large numbers of people?

If meaningful learning is reflective, constructive, and self-regulated (Herman, 1992, p. 75), the assessment paradigm required to measure the presence and extent of meaningful learning in students must also be reflective, constructive, and self-regulated. As dynamically different as these descriptors are when compared to descriptors of traditional assessments, there are some important descriptors common to both traditional and innovative assessments. These descriptors have to do with what makes an assessment defensible as an indicator of behavior. It is the development of performance-based assessments to support meaningful learning that provides the focus for Chapters 3 through 5.

Chapter 3
Why Use Performance-Based Assessment in the Classroom?

This chapter presents the linkages between innovative assessment strategies and the teaching-learning experience. Specifically emphasized is the relationship between the structure of performance-based assessment and the major themes of the NCTM Curriculum and Evaluation Standards and Science for All Americans.

Though no one is predicting the complete demise of standardized norm-referenced or criterion-referenced testing, it is certainly clear that the popular mood of the moment is towards testing situations that are innovative and allow students to demonstrate what they can really *do*, not just what they choose to "bubble in." This difference is significant in terms of the perceived value and the actual value of these assessments. Specifically, assessments that involve the performance of tasks tend to be valued in their own right, whereas multiple-choice tests have value primarily as indicators of performance in the natural setting (e.g., Linn, Baker, and Dunbar, 1991). So, innovative assessments are explicitly linked to instruction, whereas traditional assessments are less directly connected to the classroom.

Innovative assessment practices are heralded as an important key to educational reform by both critics and fans of the American educational establishment. As Harvard's Performance Collaborative for Education (PACE) project emphasizes, changing assessment practices is one of six crucial levers for realizing the vision of a thinking generation:

- Changing Assessment Practices
- Constructing Support Systems for Learning
- Changing School Structures
- Teacher Training
- Building Administrative Support
- Creating Partnerships with Families

(PACE, personal communication, 1992)

Rather than viewing changing assessment practices as separate from these other levers of change, consider the above list as all being about assessment. For example, support systems, school structures, teacher training, administrative support, and partnerships with families can all legitimately be considered as components of *changing assessment practices*. In fact, linkages among the other five levers are essential for the successful implementation of performance-based assessment in the classroom.

Assessment innovations depend upon a multitude of factors, such as: (a) teacher training in performance-based assessment, (b) administrators willing to

take risks in the area of assessment, (c) support systems for students so that they experience the kind of reflective, constructive, and self-regulated learning that is being measured, (d) changes in school structure that permit the use of assessments based on interdisciplinary content and collaboration, and (e) families willing to accept descriptive reports related to important educational themes rather than scores on discrete, chunkable instructional objectives.

If assessments are intended to support education reform and if they are intended to support a thinking curriculum—one composed of hard content and requiring complex cognitive processing from students—it is not sufficient to change the way assessments look. For example, it is a fairly simple task to take a question modeled on those found in traditional multiple-choice tests and to replace the four or five options (answer choices) with blank lines upon which students write the answers. This change in format does not change the assessment in a meaningful way. Consider the following multiple-choice test question:

The third-grade students in Mr. Stewart's class at Smoke Rise Elementary School conducted a survey to find out the kind of ice cream that they liked the most. Each "X" represents one student.

Flavor Choices

Vanilla	XXXXXXXX
Chocolate	XXXXXX
Peppermint	XXXXX
Strawberry	XXX

1. Which flavor is most popular?

(A) Vanilla*
(B) Chocolate
(C) Peppermint
(D) Strawberry

* Correct Answer

Now consider a revision that requires the student to construct a response:

The third-grade students in Mr. Stewart's class at Smoke Rise Elementary School conducted a survey to find out the kind of ice cream that they liked. Each "X" represents one student.

Flavor Choices

Vanilla	XXXXXXXXXXXX
Chocolate	XXXXXX
Peppermint	XXXXXXXXX
Strawberry	XXX

Which flavor is most popular?

_____

Superficially, the format of the question has changed. In addition, a constructed response is probably more difficult for the students because the options (answer choices) are not presented as part of the question to prompt, to structure, to suggest. Of course, guessing takes on a different character when the student must guess from internal knowledge rather than from supplied choices. But the underlying behavior of interest is still relatively discrete and low level.

In order to use performance-based assessment in a meaningful way, the focus, or underlying behaviors of interest, must change; the nature of the evidence about those underlying behaviors must change, and the way in which the information is used must change. Consider an extension of the question above:

The students in Mr. Stewart's third-grade class at Smoke Rise Elementary School conducted a survey to find out the kind of ice cream that they liked the most. Find out which flavor of ice cream is the class favorite. Document your process and display the information. You may use words, pictures, graphs, or any other method you think is easy to understand. After you have chosen a way to display this information, write about why you chose this kind of display.

The third grade students at Smoke Rise Elementary School conducted a survey to find out the kind of ice cream that they liked. Find out which flavor of ice cream is the class favorite. Document your process and display the information. You may use words, pictures, graphs, or any other way you think is most easy to understand. After you have chosen a way to display this information, write about why you made this choice.

FLAVOR CHOICES

Chocolate	Vanilla	Strawberry
Peppermint	Peppermint	Peppermint
Vanilla	Vanilla	Vanilla
Vanilla	Chocolate	Peppermint
Chocolate	Vanilla	Chocolate
Vanilla	Strawberry	Peppermint
Peppermint	Peppermint	Chocolate
Vanilla	Vanilla	Peppermint
Chocolate	Strawberry	Vanilla
Vanilla	Vanilla	Peppermint

RESPONSE SHEET

Display the information below

Why did you choose this display?

This extension of the initial question broadens the challenge of the assessment activity. The progression is from a discrete, chunkable stimulus with discrete, supplied responses, to unspecified but still discrete responses, to self-selected stimuli with open responses.

These examples, as well as those in Chapter 2, hint at the continuum as one moves away from traditional assessments to innovative, performance-based assessments. This continuum ranges from strong constraints and test developer control to high flexibility and student control. In deciding how assessment can support a thinking curriculum in your classroom, it is important to understand how variations in control affect the linkage between instruction and assessment. The movement toward greater flexibility and stronger student control of assessment is very much in keeping with recommendations for improving instruction in science and mathematics.

The Content Shift in Mathematics and Science Education

Remember that Porter, Archbald, and Tyree (1991) define hard content as:

> ...not just the facts and skills of academic work, but understanding concepts and the interrelationships that give meaning and utility to the facts and skills. (p. 11)

Indeed, *hard* content does not mean difficult. Hard content means important content, valuable ideas, principles, and knowledge without which there cannot be higher-order thinking. Another way of thinking about hard content is to think in terms of fundamental, pervasive, and essential elements of a content domain without which no student can be considered competent.

The National Council of Teachers of Mathematics (NCTM) defines content standards for the mathematics education community from a perspective consistent with Porter, Archbald, and Tyree's definition of hard content. In the NCTM *Curriculum and Evaluation Standards for School Mathematics* "a vision is given of what the mathematics curriculum should include in terms of content priority and emphasis" (1989, p. v). With the *Agenda for Action* (NCTM, 1980), mathematics educators initiated discussion that led to the *Standards* document which emerged in draft form in 1987. Two years later, the final document was published, and it stands as a powerful model for other content areas.

The NCTM *Standards* are presented as "one facet of the mathematics education community's response to the call for reform in the teaching and learning of mathematics.... Inherent in this document is a consensus that all students need to learn more, and often different, mathematics and that instruction in mathematics must be significantly revised" (p. 1). Toward that end, the document proposes curriculum goals for school mathematics.

Five general goals are that all K-12 students:

- learn to value mathematics
- become confident in their ability to do mathematics
- become mathematical problem solvers
- learn to communicate mathematically
- learn to reason mathematically

(NCTM, 1989, p. 5)

These five goals are the big ideas underlying mathematics education. It is these big ideas that provide the assessment developer with a useful and important focus.

The *Standards* go on to explain the kinds of behaviors in which students should engage as they tackle these big ideas.

> These goals imply that students should be exposed to numerous and varied interrelated experiences that encourage them to value the mathematical enterprise, to develop mathematical habits of mind, and to understand and appreciate the role of mathematics in human affairs; that they should be encouraged to explore, to guess, and even to make and correct errors so that they gain confidence in their ability to solve complex problems; that they should read, write, and discuss mathematics; and that they should conjecture, test, and build arguments about a conjecture's validity.
>
> (NCTM, 1989, p. 5)

Thus, evidence that students have the big ideas includes demonstration of mathematical habits of mind, exploration, self-regulation through error correction, reading, writing, and talking, even arguing, about mathematics.

The specific hard content elements are then detailed in the *Standards* for each of three grade-level groups (K-4, 5-8, and 9-12). Within each grade-level group are both topics and processes that should define the mathematics education experience for all students.

For the K-4 grade-level group, the topics and processes are:

Estimation	Mathematics as Problem Solving
Number Sense and Numeration	Mathematics as Communication
Concepts of Whole Number Operations	Mathematics as Reasoning
Whole Number Computation	Mathematical Connections
Geometry and Spatial Sense	
Measurement	
Statistics and Probability	
Fractions and Decimals	
Patterns and Relationships	

(NCTM, 1989, p. 15)

For the 5-8 grade-level group, the topics and processes are:

Number and Number Relations	Mathematics as Problem Solving
Number Systems and Number Theory	Mathematics as Communication
Computation and Estimation	Mathematics as Reasoning
Patterns and Functions	Mathematical Connections
Algebra	
Statistics	
Probability	
Geometry	
Measurement	

(NCTM, 1989, p. 65)

For the 9-12 grade-level group, the topics and processes are:

Algebra	Mathematics as Problem Solving
Functions	Mathematics as Communication
Geometry From a Synthetic Perspective	Mathematics as Reasoning
	Mathematical Connections
Geometry From an Algebraic Perspective	
Trigonometry	
Statistics	
Probability	
Discrete Mathematics	
Conceptual Underpinnings of Calculus	
Mathematical Structure	

(NCTM, 1989 p. 123)

The structure of the NCTM *Standards* is important in this discussion of performance-based assessment because it does not address discrete, chunkable instructional objectives typically found in state curriculum frameworks, in system scope-and-sequence documents, or in textbooks. Instead the themes in the *Standards* take the form of valued outcomes in support of the assessment paradigm shift described in Chapter 2.

Although not structured at the same level of detail as the NCTM *Standards*, *Science for All Americans* (AAAS, 1989) is another important document in education reform. Written more as a philosophy of science education than a curricular framework, *Science for All Americans* still articulates themes similar to those in the *Standards*.

Science for All Americans is the sixth in a series of reports that recommend a curricular framework in support of scientific literacy. These documents represent the first phase of Project 2061, which has established a conceptual base for science education reform by identifying the knowledge, skills, and habits of mind essential for all young people.

Recommendations that address the basic dimensions of scientific literacy in *Science for All Americans* are:

- Being familiar with the natural world and respecting its unity
- Understanding some of the key concepts and principles of science
- Being aware of some of the important ways in which mathematics, technology, and the sciences depend upon one another
- Knowing that science, mathematics, and technology are human enterprises and knowing what that implies about their strengths and limitations
- Being able to use scientific knowledge and ways of thinking for individual and social purposes.

(Rutherford & Ahlgren, 1990, p. x)

In further support of these big ideas are discussions of the *scientific endeavor*, *scientific views of the world*, *perspectives on science*, and *scientific habits of mind*. In each of these discussions are more focused examples of how science education should be structured in support of scientific literacy. For example, under the discussion on the scientific endeavor is found:

> The various natural and social sciences differ from each other somewhat in subject matter and technique, yet they share certain values, philosophical views about knowledge, and ways of learning about the world. All of the sciences presume that the things and events in the universe occur in consistent patterns that are comprehensible through careful and systematic study. Although they all aim at producing verifiable knowledge, some of them claim to produce knowledge that is absolutely true and beyond change.
>
> (AAAS, 1989, p. 5)

Under the discussion on scientific views of the world is found:

> Biological evolution as a concept based on extensive geological and molecular evidence, as an explanation for the diversity and similarity of life forms, and as a central organizing principle for all of biology. (AAAS, 1989, p. 7)

Under perspectives on science is found:

> An understanding of a few thematic ideas that have proven to be especially useful in thinking about how things work. These include the idea of systems as a unified whole in which each part is understandable only in relation to other parts; of models as physical devices, drawings, equations, computer programs, or mental images that suggest how things work or might work; of stability and change in systems; and of the effects of scale on the behavior of objects and systems. (AAAS, 1989, p. 9)

Under scientific habits of mind is found:

> Communication skills, including the ability to express basic ideas, instructions, and information clearly both orally and in writing, to organize information in tables and simple graphs, and to draw rough diagrams. Communicating effectively also includes the ability to read and comprehend science and technology news as presented in the popular print and broadcast media, as well as general reading skills. (AAAS, 1989, p. 10)

Just as is evident in the NCTM *Standards*, the big ideas in *Science for All Americans* include demonstration of analytical habits of mind, observation and exploration, and self-regulation through reading, writing, and talking.

Throughout both the *Standards* and *Science for All Americans* is a clear philosophical statement that students must construct their own understanding of important principles and knowledge, that they must use self-regulation for error correction or critical analysis, and that they must reflect upon their own work and the work of others in order to see connections, interrelationships, and the broad role of science and mathematics in the real world. It is this shared perspective across disciplines that speaks eloquently to the use of performance-based assessments as appropriate for evaluating these big ideas.

The measurement community is just beginning to understand fully the time required to create, design, develop, try out, refine, and implement a performance-based assessment activity. In many ways, developing performance-based assessments is much more difficult than creating multiple-choice tests. So, if time, talent, and resources are limited, either because they are being used for other purposes, such as teaching, or because they simply are not readily available, it is critical that performance-based assessments be used in areas of maximum "pay-off," that is, in assessing the big ideas that are rarely measured by traditional multiple-choice assessments.

With this foundation in the big ideas—important outcomes of education from the perspectives of mathematics and science educators—it is appropriate to consider now the structure of performance-based assessments.

The Elements of Performance-Based Assessments

Rethinking assessment so that instruction and assessment are inextricably linked is a challenge. Yet without the interdependency, performance-based assessment will not support educational reform, nor will it support a curriculum that encourages the development of students who think analytically, who reason, who question—in short, who use their minds well.

According to Gregory Anrig, past president of Educational Testing Service, the elements of the new generation of assessment:

- link assessment and instruction
- are individualized and adapted to the student's abilities
- provide more useful information
- mirror real-life skills.

(Anrig, 1991, p. 1).

Two of these four elements are common to traditional assessment modes as well. For example, test developers and educators have consistently sought tests that link assessment and instruction. The clamor for more useful information and more meaningful reports for students and parents, has also been consistent throughout the era of multiple-choice tests.

The other two elements are not common, however. Test developers have seldom argued that traditional multiple-choice tests are individualized and adapted to the student's abilities or that they mirror real-life skills. Some few tests may be designed or marketed for different categories of students (i.e., visual learners, auditory learners, etc.) but the same multiple-choice test is generally not designed or marketed to serve the needs of diverse populations. Similarly, a multiple-choice test may be touted as predicting real-life or on-the-job performance, but the test itself is not likely to mirror real life.

Thus, Anrig is calling for notable changes in assessment. These changes are most suited for use in the classroom, where linkages to instruction are most useful, where the degree of adaptability to support learning is most needed and most readily handled, where rapidly retrievable information is desired by teachers and students, and where the classroom itself can become a microcosm of the real world.

These elements provide a useful framework for the redefinition of assessment. However, some specifics are essential before embarking on the development of performance-based assessments.

In order to create assessments to support the kind of instruction that empowers students to use their minds well, an assessment model must be designed that reflects ideal instruction and emphasizes higher-order thinking skills. So, the context for assessment must reflect integrated skills, realistic situations, and dilemmas for which there are multiple legitimate strategies and multiple correct solutions (see Educational Testing Service, 1991). These assessments seek to measure directly the student's ability to perform in the subject area (Willis, 1990).

Before listing some of the essential elements that should be either incorporated into each performance-based assessment or intentionally and reasonably excluded, it is important to precede this discussion with a caution: The work in performance-based assessment is evolving. As more assessments are developed and field-tested, as more data are collected, and as more information about implementation and use is collected, the perspective on these essential elements will undoubtedly change. Consider these elements with curiosity and a healthy degree of skepticism. Ten years from now, it is likely some will remain, others will have been deleted, and still others will have been modified significantly.

Whether these characteristics are typically found in multiple-choice tests is a question to be answered through close scrutiny of available tests. Current rhetoric suggests that typical multiple-choice tests do not embody these characteristics in sufficient strength to make them effective change agents within the classroom. But it must not be assumed that because assessment is called performance-based, it will automatically have these important characteristics. Similarly, it should not be assumed that because an assessment is multiple-choice, it is *ipso facto* not the most efficient and accurate way to describe a learning outcome. Perhaps the best caveat is to beware of labels. Check the ingredients!

As performance-based assessments are used to support change in mathematics and science education, it is important that these assessments reinforce those characteristics of high-quality instruction that are most likely to encourage complex cognitive behaviors. Toward that end, performance-based assessments should:

- reflect ideal instructional practices
- incorporate production tasks
- involve the teacher as participant/observer
- require collaboration
- prompt investigation
- be motivational and promote natural curiosity
- facilitate use of multiple strategies
- yield multiple solutions
- emphasize big ideas
- incorporate multiple goals
- integrate knowledge and process
- be relevant and topically current
- reflect an appropriate level of difficulty
- be feasible
- be cost effective
- tap higher-order thinking skills

Most of these descriptors are familiar. Teachers know precisely which assessment topics have relevance and currency for their students; it is no surprise to teachers that multiple-choice tests often do not interest their students.

Teachers know what an appropriate level of difficulty means for their students; they know which students can handle difficult material easily and which students can handle only the most basic material.

Teachers know what multiple goals are; they know when a solution to a single question requires knowledge and processes from multiple domains, either across or within disciplines.

But when educators talk about higher-order thinking, they are often talking about different behaviors. Some may rely on Bloom's *Taxonomy* (1956) and define higher-order thinking as *application* and above. Others may say that everything that is not a function of rote memorization is higher-order thinking. Still others may say that the term *process skills* is the appropriate definition for higher-order thinking skills.

Resnick (1987) suggests that higher-order thinking:

- is nonalgorithmic,
- tends to be complex,
- often yields multiple solutions,
- involves nuanced judgment and interpretation,
- involves the application of multiple criteria,
- often involves uncertainty,
- involves self-regulation of the thinking process,
- involves imposing meaning and finding structure,
- is effortful.

If emerging learning theory is correct, higher-order thinking is best described as a web rather than a hierarchy (see Figure 1). As students construct meaning based on their prior knowledge and the context in which they are operating, each student is likely to process information in a unique and individual way. So, as each student formulates questions in search of solutions, each is likely to move through the web in a unique way, approaching a solution from the idiosyncratic perspective of his/her own individuality. The web of information processes may then be defined in terms of questions posed as the student observes, makes predictions, investigates, reevaluates, attempts, observes, makes additional predictions, observes, and so on. In the web, the points of intersection represent the knowledge essential to solving the problem. The connecting strands of the web represent the processes or strategies used to move through the information (knowledge).

As each student traverses the web, with some students being more efficient than others, progress from the identification of a problem to its solution is most unlikely to be linear or hierarchical. So, it seems quite evident that linear, hierarchical models like Bloom's *Taxonomy are* not consistent with the way students learn. On the other hand, Resnick's list of characteristics allows the process of *using* higher-order thinking skills to be described from multiple and complex perspectives consistent with individual differences in thinking and learning.

How then, does Resnick's list support the big ideas of the reform movement in science and mathematics? It seems quite clear that throughout the list, words like *estimating*, *hypothesizing*, *investigating*, and *observing* are synonyms for many of the words or phrases, thus the big ideas, found in both *Science for All Americans* and the NCTM *Standards*.

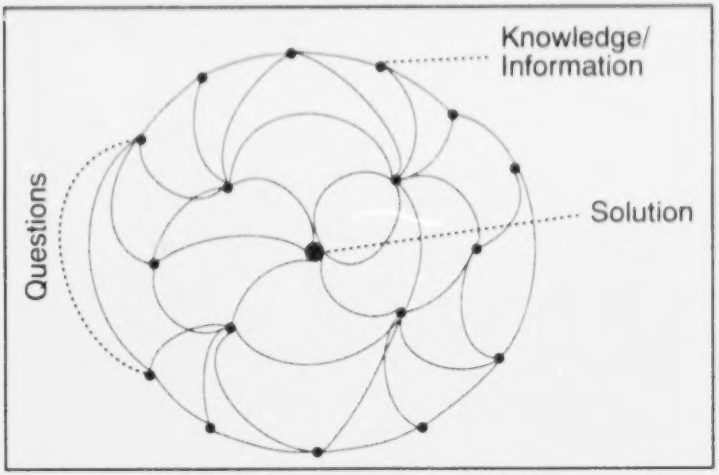


Figure 1. Information Processing Web.

Chapter 4
How Can Performance-Based Assessment Really Work?

This chapter presents a structure for designing performance-based assessments for classroom use. Guidelines for creating a task that supports specific purposes and documents big ideas are given. "The Soda Task," developed by the Connecticut State Department of Education, is used for illustration.

Assessments, whether traditional or innovative, can serve many purposes. Assessments can be used to sort, classify, affirm, diagnose, prescribe, or characterize when applied to individuals. Assessments can also provide a basis for evaluation when applied to programs or curricula. But, before beginning development of any type of test or assessment, the precise purpose for the instrument must be clearly articulated. The first question that must be asked and answered is *What purpose will your test (assessment) serve?* All of the design and implementation questions posed in this chapter require answers that reflect the purpose to be served.

Because this book is intended for use by classroom teachers and because it is in the classroom that the most important changes will result from performance-based assessment, this chapter assumes the answer to the question of purpose is *to inform instruction*. If your purpose is different, you must still answer the questions that follow, but your responses may be quite different. As you progress through these decision points, it is important to keep in mind that the focus for this discussion is an instrument whose purpose is to inform instruction as it measures the academic growth of students.

Focusing the Assessment Development Process

Once the purpose of an assessment has been identified, the next decision points are:

- What are you trying to *describe*?
- What must be *documented*?
- What should the assessment *model*?
- Whom (and how) are you trying to *inform*?

These four questions define the paradigm for performance-based assessment. Answers to these four questions establish parameters and rules that govern the design, implementation, and interpretation of performance-based assessment.

Suppose the following: Teacher A wants to design an assessment that will identify the extent to which students can apply the scientific method to a problem set outside the classroom (i.e., a real-life problem). This purpose provides a focus

for the development of a performance-based assessment. What defines and determines the specifics of that assessment, however, are the responses to these four questions:

- What is Teacher A trying to *describe*?
- What must Teacher A *document*?
- What should Teacher A's assessment *model*?
- Whom (and how) is Teacher A trying to *inform*?

In terms of Teacher A's focus (the appropriate and effective use of the scientific method on a problem set outside the classroom), what needs to be described? Clearly, the process by which the student decides upon an analytic framework to solve the problem, the concise articulation of the problem, and the strategies for implementing the scientific method must be described. The documentation must include permanent or archival responses to the assessment stimulus. And, because Teacher A uses cooperative learning in the conduct of science education, the assessment should model collaboration. Results of the assessment will influence Teacher A's instructional practice and inform students of their academic progress.

The Look of Assessment

A fundamental goal cited by Lindquist over 40 years ago, to make tests as nearly equivalent to the desired learning outcomes as efficiency and economy permit, does not change with the paradigm shift from traditional to innovative assessment. Instead, the possibility of realizing Lindquist's goal has become more likely as assessment tools and desired behaviors become indistinguishable. The boundaries between instructional or learning activities and assessments become blurred (see Baron, 1990, 1991), and performance-based assessments emerge looking very much like instructional activities.

A superb example of this blurring between assessment and instruction is a performance-based assessment activity called "The Soda Task" (Connecticut Department of Education, 1989), an assessment that takes the form of an experiment:

> **The Soda Task**—You will be given two samples of soda, one regular soda containing sugar and the other one diet soda containing an artificial sweetener. Your task is to identify each sample as diet or regular based on your knowledge of physics, chemistry, and/or biology. As in any experiment, you are not allowed to taste any of the samples.

Background information is provided, and students are guided to brainstorm with their peers to design and conduct an experiment to distinguish between the two samples and then report their findings (see Appendix D for a complete description of the task).

A good teacher of science, especially at the high school level, may review "The Soda Task" and say, "That's no different from the way I teach every day." Indeed, in many ways this assessment task does represent good instructional practice. It is hands-on science. It is scientific inquiry.

In order for "The Soda Task" to be classified as an assessment, it must meet the requirements of Cronbach's (1970) definition of test:

> A *test* is a systematic procedure for observing behavior and describing it with a numerical scale or category system. (p. 26)

What is it that identifies "The Soda Task" as an assessment activity rather than an instructional activity? Clearly the task focuses on student behavior, and that behavior can be communicated via summary scores or category descriptors. The real question is whether the task provides for systematic observation.

To answer this question, it is helpful to think about *systematic* in a slightly unusual way. When used to describe traditional multiple-choice tests, systematic typically means *standardized*, implying that every examinee experiences precisely the same conditions (including the amount of time, stimuli materials, and response format). If this interpretation is applied to "The Soda Task," this innovative assessment activity does not qualify as systematic. In "The Soda Task," students are allowed to collaborate. Because the group dynamics and skill levels of the students are likely to vary considerably from group to group, the experiences within each group will typically be different, so the activity would traditionally be considered *nonstandardized* and, therefore, unsystematic. In performance-based assessment, however, there is often the desire to make the experience nonstandardized, just as authentic, real-world experiences are. Thus, a new way of thinking about standardization and systematization begins to emerge, not as absolutely precise characteristics and limitations on test-taking behavior, but as boundaries and parameters within which variation occurs as a natural consequence of diversity among the students participating in the assessment activity. Using this interpretation of systematic, "The Soda Task" does indeed provide for systematic observation and can therefore be legitimately used as an assessment activity.

What does "The Soda Task" activity describe? What does it document? What does it model? Whom and how does it inform? An analysis of "The Soda Task" from the perspective of these four questions is fairly straightforward.

◆ *What does "The Soda Task" describe?*

In an earlier version of "The Soda Task," the behaviors targeted were listed as the following:

Students should be able to:

- Identify and apply physical and/or chemical properties for the purpose of identification;

- Use and make measurements using appropriate units;
- Formulate predictions based on prior knowledge;
- Identify information and steps needed to solve problems;
- Test predictions;
- Gather data pertinent to a problem;
- Make inferences based on pertinent data;
- Draw reasonable conclusions and defend them rationally;
- Communicate the strategies and outcomes of a study through written means;
- Communicate the strategies and outcomes of a study orally;
- Work cooperatively in a group.

(Baron, Forgione, Rindone, Kruglanski, & Davey, 1989, p. 2)

Each of these objectives clearly identifies the behaviors of interest to the assessment developer and, by implication, to the assessment user. These objectives focus on scientific habits of mind (i.e., using information in a systematic way to answer important questions and to communicate this information in a variety of ways, both written and oral). Although framed as a science assessment activity, "The Soda Task" not only supports the concept of scientific literacy advocated in *Science for All Americans*, it also taps the content area of measurement and the processes of problem solving, reasoning, communication, and connections promoted in the NCTM *Standards*. Moreover, this task incorporates, by its very design, a focus on cooperative social behavior.

◆ *What does "The Soda Task" document?*

The instructions to the student (see Appendix D) detail the specific documentation strategies used. In Part I, each student is to list possible ways of identifying differences between the two sodas. In Part II, steps 2 - 5 focus on thinking and communicating, rather than doing. As students brainstorm and carry out experiments, what method of documentation is available to capture evidence of the initial assessment objectives? Is this method available to the teacher or to the parent or to the student? Steps 1, 2, 3, and 4 require a group written product of some known dimension. Step 5 requires preparation and delivery of an oral presentation. Step 6 requires a final and presumably written group product. Part III provides an opportunity for each student to produce some individual documentation.

The amount of detail in the documentation depends on the complexity of the experiment and upon the individual student's understanding of the thoroughness required. The teacher does have, however, the opportunity to intervene and/or provide feedback when shown the experimental plan in Step 3 of Part II.

◆ *What does "The Soda Task" model?*

This assessment activity models scientific investigation in the classic sense. But it does much more than that. First of all, "The Soda Task" is an interesting

blend of individual and small-group work. Secondly, it is complex in tapping multiple goals and facilitating multiple strategies. It is cognitively complex, as well, in requiring higher-order thinking skills. Fourth, it requires that each student cooperate with others in the work group. And, it models for teachers how ideal instruction should look.

- *Whom does "The Soda Task" inform?*

A review of the scoring guide in Appendix D suggests that the data collected through this assessment are structured to correlate with the 11 behavioral objectives listed above. The specific audiences who could be informed through this scoring guide are determined by the relative literacy of these audiences. Certainly, to someone familiar with science education, scores in these objective categories would make a great deal of sense. However, it is fair to ask whether this information would be meaningful to audiences with less background in science education. How different from a traditional test score can an assessment report be and still be interpretable by a lay audience?

Properties of Performance-Based Assessment

If an activity "passes" these critical questions, it is appropriate to consider it an assessment and not just an instructional activity. The next critical questions parallel the elements listed at the end of Chapter 3. They focus on the properties of innovation, the majority of which should be present if the assessment is to be considered high-quality and performance-based.

- *Does the assessment reflect ideal instructional practice?*

In reviewing an assessment to determine if it models ideal instructional practice, it is important to conduct that review in the context of current understanding of how individuals learn. This means that the assessment developer must keep abreast of research in cognition.

Research has demonstrated that hands-on experience with manipulatives is the most effective way for children to internalize complex cognitive behaviors (Slavin, 1991). Research has also demonstrated that instruction that emphasizes inquiry facilitates higher-order thinking skills in learners. Furthermore, research has demonstrated that cooperative learning has major social, as well as academic, benefits for students.

These three facets of instructional practice suggest the ways in which assessment should reflect ideal instructional practice. Performance-based assessment should require hands-on experience with manipulatives, should include inquiry-based stimuli, and should encourage cooperative learning among students.

- *Does the assessment incorporate production tasks?*

The result of hands-on activities is typically a production task of some sort. These production tasks may be written or spoken performances, constructed models, visual arts, or performing arts. But, the production task itself need not be the only aspect of the assessment for which performance is judged. For example, the process through which the product is developed can also be a focal point for the assessment. This is often the case in situations where the solution to a problem is less enlightening than the strategies followed to solve the problem. For instance, in "The Soda Task" the particular experimental framework that students use to solve the problem is much less enlightening than the strategies and processes employed by the students in their search for that solution.

- *Does the assessment involve the teacher as participant/observer?*

Because this discussion is about assessment used to inform instruction, it is appropriate to think of the teacher as participant as well as observer. Of course, it would be ludicrous to consider the teacher as anything other than the reserved and aloof test administrator in a high-stakes assessment (i.e., a test used for the purpose of passing, failing, or promoting individuals). But in most classroom assessments, the assessment activity is a learning experience. If the individual or group of individuals is stumped on how to move ahead through the problem, it makes considerable sense for the teacher to act as a "nudge" to facilitate that movement.

Clearly, the extent to which the teacher nudges, and the type of nudging, need to be noted in the performance documentation. If the nudge is content related, for example, that documentation is important in understanding the breadth and depth of content knowledge demonstrated by the students. If the nudge is related to processes, the interpretation of the nudge must be relative to those processes. If the nudges have to do with behavior problems or group dynamics, then those nudges must be interpreted in terms of the outcomes related to the students' ability to collaborate.

Even while participating in the activity, the teacher must always be alert as an observer. This standard applies whether the teacher is engaged in performance-based assessment or simply being an effective and responsive teacher.

- *Does the assessment require collaboration?*

Successful and effective collaboration is a major societal and work-place goal as well as a valued outcome within the context of schooling. Children, like adults, must practice collaboration across the various groupings of their social microcosm. Just as diversity and inclusion are paramount societal goals, so too must they be underlying characteristics of performance-based assessment used to inform instruction. The key phrase here is *to inform instruction* because in no situation should collaboration be considered in high-stakes assessment programs from

which promotion, retention, selection, and hiring decisions are made. Until there is a way to quantify or qualify fairly the individual contributions revealed through collaborative assessment, collaboration merely clouds the scoring picture.

- *Does the assessment prompt investigation?*

To say that performance-based assessment must allow investigation is an understatement. Particularly in science and mathematics, it is more appropriate to say that performance-based assessment should require investigation. Performance-based assessment in science and mathematics must capture the inquisitive mind. It must employ scientific reasoning processes. It must trigger hypothesizing and observation. It must encourage trial and error or estimation.

- *Is the assessment motivational and does it promote natural curiosity?*

If performance-based assessment is interesting and engaging, it will be motivational. Students will want to complete the assessment activity, not because they must finish within the time allowed, but because something in the task intrigues them. It is the hook of intrigue that qualifies a performance-based assessment as motivational.

Performance-based assessment should be fun for students. Ideally, it should be fun for *all* students. Realistically, performance-based assessment will be fun for most students, and that in itself is motivational. Students should react to the closing of a performance-based assessment task as the beginning of continued investigation.

- *Does the assessment facilitate the use of multiple strategies?*

In life, there is seldom one and only one right way to move from problem specification to problem solution. So, too, in performance-based assessment there should be more than one productive and judicious way to move from the statement of the problem to the solution.

- *Does the assessment yield multiple solutions?*

Richard Lesh (personal communication, 1992) is quite adamant in stipulating that a performance-based assessment task must lend itself to multiple solutions. The point is that the problem to be solved must be sufficiently complex that no one single solution is always right.

- *Does the assessment emphasize big ideas?*

In "The Soda Task" assessment activity, the problem is to design an experiment that provides information for decision-making. This process provides evidence of students' capacity for using scientific ways of thinking for individual

and social purposes. Key concepts and principles of science are tapped in this assessment as well. The extent to which specific habits of mind are tapped is largely a function of individual examinees' responses to the task.

- *Does the assessment incorporate multiple goals?*

The presence of multiple goals in "The Soda Task" assessment is quite evident. A review of the 11 objectives listed earlier in this chapter indicates that there are process goals, content goals, and social goals.

- *Does the assessment integrate knowledge and process?*

As many authors have pointed out, traditional tests in both science and mathematics focus primarily (73–96%) on low-level thinking skills (Madaus et al., 1992, p. 3). Multiple-choice tests tend to contain items that evoke only knowledge-level information from examinees. Thus, one seldom, if ever, gets information about process from these tests. Within the measurement field, one of the major advantages of performance-based assessment is considered to be the possibility of detecting, by the nature of the tasks, both process and product.

- *Is the assessment relevant and topically current?*

It is probably unreasonable to expect that any single performance-based assessment will be relevant and current for each examinee. There is a high probability, however, that high school students would be interested enough in soda, for example, to consider "The Soda Task" relevant. Another important feature is that this task would undoubtedly be judged as being racially and ethnically fair.

- *Does the assessment reflect an appropriate level of difficulty?*

Like any traditional assessment, performance-based assessments begin as design ideas. This design idea typically emerges from an instructional need or experience. If the instructional idea is targeted to the appropriate grade level, the assessment is likely to be appropriately targeted as well. However, as in traditional assessments, performance-based assessments must be field tested with real students in order to determine empirically if the assessment is appropriately positioned. It is this empirical reality check that ultimately determines whether or not the assessment reflects an appropriate level of difficulty.

- *Is the assessment feasible?*

What is feasible in one classroom situation may not be in another. Certainly the administrative ease of the performance-based assessment contributes to its feasibility. But so also does the opportunity, real cost, and complexity of the

assessment. The classroom teacher must be the ultimate arbiter of what assessment is feasible to administer in his or her classroom.

- *Is the assessment cost effective?*

Although "The Soda Task" is relatively feasible and practical as reported by Baron et al. (1989, p. 17), its true cost effectiveness must be judged relative to the utility of the information derived. Here one could question whether or not the assessment itself provides sufficiently rich information to warrant the expense of design, implementation, scoring, and reporting. One might argue, for example, that because this assessment is so like the traditional instructional format found in a high school chemistry class (experimentation and lab reports) that the addition of this structured assessment is not likely to provide enough new information to the already available data base to warrant the expense. On the other hand, if the development of the assessment serves to move teachers toward systematic observation of complex cognitive behaviors (multiple goals and multiple strategies), it may be invaluable as both an efficient and effective activity.

- *Does the assessment tap higher-order thinking skills?*

In assessing the extent to which a given assessment taps higher-order thinking skills, Resnick's list may be useful (see end of Chapter 3). Certainly, from this perspective, "The Soda Task" indisputably taps higher-order thinking skills.

To summarize, a high quality performance-based assessment should satisfy the vast majority of criteria in the *Assessment Rating Form* at the top of the next page. A review of "The Soda Task" is included for purposes of illustration.

If performance-based assessments are designed from the perspective of these essential elements, they are likely to reveal important and dynamic aspects of thinking that are seldom described or documented for either teachers or students. Furthermore, if performance-based assessments are designed to address the majority of these elements, the activities will indeed model exemplary, student-centered instruction. Finally, if performance-based assessments are designed from the perspective of this *Assessment Rating Form*, the potential for providing information that will be useful for informing instruction is significant. Just as "The Soda Task" provides an excellent example of worthwhile performance-based assessment for the classroom, so too will other assessments having these characteristics.

Organizing the Assessment Development Process

Organizing the development of a performance-based assessment can be a considerable challenge. Having identified the essential conceptual elements that need to be addressed, it is useful to have an organizational framework, as well. At the bottom of the next page is one such basic framework.

Assessment Rating Form

Characteristics	No	Somewhat	Very Much
• Is the assessment equivalent to criterion behavior?		X	
Does the assessment satisfy the definition of test?			
• Is it systematic?	X		
• Does it focus on behavior?		X	
• Does it yield numerical or category scores?		X	
Does the assessment have the important properties of performance-based assessment?			
• Reflects ideal instructional practices?		X	
• Incorporates production tasks?		X	
• Involves the teacher as participant/observer?	X		
• Requires collaboration?		X	
• Prompts investigation?		X	
• Is motivational and promotes natural curiosity?	X		
• Facilitates use of multiple strategies?		X	
• Yields multiple solutions?		X	
• Emphasizes big ideas?		X	
• Incorporates multiple goals?		X	
• Integrates knowledge and process?		X	
• Is relevant and topically current?	X		
• Reflects an appropriate level of difficulty?		X	
• Is feasible?		X	
• Is cost effective?	?		
• Taps higher-order thinking skills?		X	

OUTLINE

PERFORMANCE-BASED ASSESSMENT ACTIVITY

Overview/Purpose:

Time Limit:

Materials:

Setup:

Directions:

Follow-up:

Scoring Rubric(s):
(possible solutions)

Use of this outline is demonstrated below as "The Soda Task" information from Appendix D is restructured into this framework:

OUTLINE

THE SODA TASK

Overview/Purpose:

The purpose of "The Soda Task" is to determine whether students can interact both with a sophisticated knowledge base of important scientific principles and with other students to reach a conclusion. The content domain is the physical, chemical, and biological properties of matter. *[Specific objectives underlying this assessment are listed at the beginning of this chapter.]*

Time Limit: 3-4 class periods

Materials:

Regular soda	Beakers	Graduated Cylinders
Diet Soda	Tripods	Heat Source
Wire Gauze	Safety Glasses	Aprons

Optional: Yeast, Benedict's Solution, and Glucose Test Strip
References: *Merck Index, CRC Handbook of Chemistry and Physics*, chemical dictionary, and chemistry textbooks

Setup:

Normal laboratory safety procedures should be followed. Students should wear safety goggles and aprons at all times. The teacher will have to prepare samples of regular and diet soda and label them A and B. The number of samples should be sufficient for each group to conduct several experiments. The teacher should provide equipment and materials to support a variety of inquiry methods. *[It would appear that laboratory stations would be the appropriate organizational scheme to be used for this activity. However, work tables around which students can sit would also be appropriate for the brainstorming and writing aspects of this assessment.]*

Directions:
[See Instructions to the Student, Parts I & II, Appendix D]
Follow-up:
[There are no explicit connections between this assessment activity and ongoing instruction. However, it is not difficult to think of immediate extensions. For example, see Instructions to the Student, Part III, Appendix D.]
Scoring Rubric:
[See Scoring Guide, Appendix D]

It should be noted that, for an activity like "The Soda Task," the descriptive materials go well beyond this basic outline. The categories indicated above constitute the minimal framework for describing a performance-based assessment activity.

This organizational framework is useful for at least two reasons. First, it provides a structure that enables the assessment developer to examine and reexamine the basic conceptual and operational requirements for the assessment. For example, it enables a reviewer to have a clear understanding of the purpose for which the assessment was developed. This, in turn, provides a vehicle for evaluating each of the design characteristics to determine whether or not it is consistent with the purpose of the assessment.

Second, it serves as a continual prompt for the assessment developer so that critical elements are not overlooked or assumed. It is indeed a dangerous adventure to develop performance-based assessment with few or no structural reminders. Thus, if one combines the structure of the organizational outline with the essential elements of the *Assessment Rating Form*, there is a reasonable probability that the development process will yield a rich and meaningful performance-based assessment activity.

Chapter 5
How Can Scoring Rubrics Communicate Complex Information?

This chapter provides an overview of the use of human judgment in scoring student performance. The process of developing scoring rubrics is described, including holistic versus analytic scoring, determining score points, and articulating content and performance standards.

Though some of the terminology may be recent, performance-based assessment actually predates multiple-choice testing by hundreds of years. Testing experts around the world have struggled with the same issues now being faced in scoring performance-based assessments. These issues are the time required for scoring, accuracy and reliability of scores, and usefulness of the scoring information.

The information gleaned from a performance-based assessment activity has limited usefulness if it cannot be communicated, aggregated, or tracked in a concise manner. Without the translation from observed behavior to numerical scale (quantification) or category system (qualification), the information is useful only to the extent that the user can retain the details accurately. Given a typical context for instructional assessment in a classroom of 25-30 students, it quickly becomes clear that asking a teacher to remember the details of each individual student's behavior or even of each collaborative team's behavior would be an unreasonable expectation. There needs to be a system for managing rich information. This system is the scoring guide, or rubric.

Scoring rubrics for performance-based assessments are essentially the scoring templates that are superimposed on performances in order to translate those performances into brief descriptors (i.e., numerical scales or category systems). Scoring rubrics in science and mathematics are heavily influenced by those that have been used in the direct assessment of writing. Wiggins, Browne, and Houston (1991) point out this interesting piece of information about the word *rubric*:

> A rubric is a set of scoring guidelines for giving scores to student work. The word derives from the Latin word for *red* and was once used to signify the directions for conducting religious services, found in the margins of liturgical books—and written in red. (p. G-10)

The "ideal" test from the perspective of teachers and administrators may be one that could be administered in about 20 minutes, is self-scoring, diagnostic, prescriptive, and comprehensive, and which provides accurate and meaningful information to students, parents, and decision makers. The "ideal" multiple-choice test does not exist, and it would be a mistake to think that performance-based assessments can accomplish all of these goals either.

The search for such a test will probably never end. In fact, this quest is full of contradictions. A test that can be administered in only 20 minutes can hardly be either diagnostic or comprehensive. An effective diagnostic assessment tool or an effective comprehensive assessment tool simply requires more time. A test that is self-scoring eliminates the possibility of evaluating individually crafted responses, often the best source of diagnostic information. In fact, the more authentic the performance-based assessment is, the less likely it is that the assessment will serve multiple purposes. It is important to remember that from what educators now know about performance-based assessment, the demands on time for development, administration, and scoring and interpreting results often exceed those for traditional assessments. Thus, the investment in performance-based assessment must focus on important knowledge, processes, and skills that are not more effectively and efficiently assessed in more economical ways.

To develop a scoring rubric, it is important to begin by revisiting the stated purpose of the assessment. This purpose will determine the level of specific detail required in the scoring rubric. If the purpose of the assessment is to identify which students have mastered a particular unit of instruction, for example, the rubric need support only two decisions: mastery, non-mastery. If, however, the purpose of the assessment is to yield diagnostic information about a student, the rubric must be sensitive to many more variations in performance than are revealed in a dichotomous scale.

Similarly, if the purpose of the assessment is to reveal relative strengths and weaknesses with respect to a broad domain of content and process, the scoring rubric must reflect that breadth and depth. If, on the other hand, the assessment is to be used to reflect relative strengths and weaknesses with respect to a narrow and focused domain of content and process, the scoring rubric must target those specifics. If the assessment is designed to be diagnostic, the assessment must be chunkable into the small building blocks of learning, with the goal being to identify deficit areas to which intervention can be applied to improve the chances of success in the next content chunk. If the goal is to determine competency or mastery, collaborative efforts may be inappropriate. After all, mastery is typically defined on individual terms. Would it be fair to determine mastery based on group performance?

The consequence of selecting one or more of these foci is not unique to innovative testing practices. It is not because educators are exploring performance-based assessments that they have to make a decision about purpose and then live with some of the implications of that decision. Just because a new term has been coined and the rhetoric for revolution is energizing, there is still the issue of fair use of test information. The development of a rubric that does not address the purpose contributes as much to misuse as does the development of an assessment task that does not address the intended purpose.

Performance-based assessments are supposed to be rich with information about how students learn and about what they have learned (i.e., the process and product of learning). However, the richer and more complex the assessment activity and the more it is sensitive to individual differences with regard to thinking

strategies, the more unlikely it is that the assessment will function as a diagnostic assessment for every student. It is highly likely that the assessment activity will be diagnostic only if the student chooses for it to be. In "The Soda Task," for example, if the assessment had not requested that the students make lists, and be prepared to explain choices, the process information would have been available only for those students for whom this documentation was a natural and normal activity. For those students with less concern or interest in documenting the steps in designing the experiments, this information would have likely been lost.

Clearly, therefore, the purpose of the test is closely tied to the documentation strategies chosen. These two elements of assessment design limit and define the scoring rubric in very practical ways.

Just as in other forms of tests, performance-based assessment must have a scoring framework or template that can be used to ascertain the correctness of different responses. In the case of performance-based assessment, *correctness* is not the same as one right answer per question. It is a continuum of right answers with characteristics and properties designed for individually scoring the performance, rather than fixed numbers or words. Because a continuum of right answers is the norm, not the exception, the structure of both the scoring guide (rubric) and the training directions for implementing the scoring guides must be scrutinized carefully.

Rubric Development

In traditional test development, the process of identifying the key or the correct response is initially performed by the item writer. Then, as the item is reviewed by various content, measurement, and editorial experts, the key is constantly scrutinized. Once agreement has been reached about the structure, wording, and key, that item is then field-tested. With the collection of data from the field tests, the item is then reevaluated, including an empirical verification of the key.

The process of developing the key for performance-based assessment is parallel to the process described above. An assessment activity (item) is developed for which a key (set of correct responses or valid characterization of correct responses) is developed. This key, or scoring rubric, is then reviewed along with the assessment task both before and after field-testing, at which time empirical feedback is available to inform revisions. The scoring rubric is, of course, more complicated than a single letter denoting one of multiple options, but its intent and use are the same. It is the standard against which student performance is judged in order to determine achievement of the examinee. The scoring rubric is the mechanism for moving from student behaviors to numerical scores or category descriptions.

In order to identify the steps that are important in the development of scoring rubrics, the history and growth of the direct assessment of writing movement provide the most useful information. The parallel between the task of producing a written product and that of completing a performance-based activity provides a

wealth of experience to draw upon in developing rubrics, training raters, and processing large numbers of product outcomes.

Stalnaker (1951) defined writing assessment as follows:

> The essay question is defined as a test item which requires a response composed by the examinee, usually in the form of one or more sentences, of a nature that no single response or pattern of responses can be listed as correct, and the accuracy and quality of which can be judged subjectively only by one skilled or informed in the subject. The most significant features of the essay question are the freedom of response allowed the examinee and the fact that not only can no single answer be listed as correct and complete, and given to clerks to check, but even an expert cannot usually classify a response as categorically right or wrong. Rather, there are different degrees of quality or merit which can be recognized. (p. 495)

In the development of scoring rubrics, it is particularly important to attend to the "freedom of response" element quoted above and discussed in Chapter 4. It is the relative unpredictability of this response that presents the greatest challenge to rubric developers; the rubric must enable raters to translate performance of varied types and at various levels to a scoring continuum in a fair and reliable way. In other words:

> The fewer the restrictions on assessment responses, the greater the reliance on human judgment for interpretation. (Jorgensen, 1991)

Holistic Versus Analytic Scoring

There are two major categories of rubric design that are used in performance-based assessment: holistic and analytic. Holistic scoring is when raters make a single, overall judgment of the quality of the response (Hogan and Mishler, 1982). Analytic scoring is when raters score each performance on specific and different elements of the task, with the combination of these elements reflecting overall performance. In each case, there are criteria for levels of performance that are decided upon by expert judges. These criteria are clearly articulated, and scorers or raters apply these criteria to each performance example. The quality control checks on scoring focus on the ability of raters or scorers to apply the criteria consistently.

In terms of "The Soda Task," note that the scoring rubric (Appendix D) indicates that there are nine distinct aspects of performance that are to be judged. Ratings on these nine can then be aggregated to form a composite score for the assessment task. This framework is analytic because it focuses the rater on specific and different elements of the task.

If, however, the scoring rubric asked only if the group were able to design an experiment to investigate the problem, the scoring would be holistic because the raters make a single, overall judgment about the quality of the response.

These differences in level of detail for reporting are a potent reminder of why the scoring rubric must be in concert with the purpose of the assessment. If, on the one hand, diagnostic information is desired, an analytic scoring rubric would be desirable, perhaps indispensable. If, on the other hand, the assessment is intended to present a comprehensive picture of performance, then holistic scoring may be all that is needed.

An example from writing may be useful at this point. Consider an assessment to determine whether students can write a persuasive letter. If this task were administered at the end of the year, a holistic scoring rubric would provide an overall assessment of students' capabilities. If the assessment were going to be used to tailor instruction in writing, an analytic scoring rubric would focus not only on the overall quality of the persuasive letter but also on the mechanics, vocabulary, syntax, tone, and organization. If the rubric used is holistic, information necessary to tailor instruction is not likely to be available without returning to the original response and evaluating it for specific elements of writing. If, however, the rubric is analytic and the elements of the analytic rubric are clearly connected to instruction, there would be no need to return to the original document. The trade-offs are between the level of detail in the information provided by the scoring rubric and the time required in scoring.

As is clear in "The Soda Task," there is a definite relationship between the underlying instructional objectives (see Chapter 4) and the elements to be formally scored (see Appendix D). It is only the second objective listed (students should be able to use and make measurements using appropriate units) that is not explicitly included in the rubric. Implicit treatment of this objective suggests that the developers believe that this skill is important in the task but does not warrant separate reporting for the test *consumers*.

By linking the scoring rubric directly to the objectives, "The Soda Task" developers provide an excellent example of how the content objectives underlying the assessment task can be used to structure the scoring and reporting process. In a sense, the analytic scoring rubric maps for the user the content standards for this assessment. If the assessment task is developed for a clearly articulated purpose, then the linkages between the behaviors of interest (objectives) and the scoring rubric should be easily identifiable and clearly reasonable.

Holistic scoring is not intended to provide detail through the scoring process. Whether holistic scoring or analytic scoring is the appropriate vehicle for use depends upon the purpose of each assessment task and its intended use. It is reasonable to expect that assessments used within the classroom for the purpose of instructional feedback to the teacher and student will be analytic rather than holistic.

Part of the decision about which type of scoring rubric to use is based upon how accurate or stable the information obtained must be. This question leads directly to the question of reliability. After almost 20 years' experience in the direct

assessment of writing, it has been estimated that interrater reliability runs .80 or above when a holistic scoring rubric is used. In analytic scoring situations, the data tend to be similar (Hogan and Mishler, 1982). The extent to which these results may generalize from the direct assessment of writing to other areas of performance-based assessment is a question yet to be answered. Each developer can contribute to this knowledge base by researching questions like these throughout the development and implementation of performance-based assessment.

Other issues may impact decisions about which approach to scoring—holistic or analytic—is better. For example, it is generally faster to make fewer decisions per assessment than to make many decisions. Thus, given certain time constraints, it may not be feasible to use analytic scoring rubrics. On the other hand, if the situation demands less global information, holistic scoring may not be adequate.

Like so many of the issues touched upon earlier in this text, there are no right or wrong decisions that can be applied across the board. Reflect upon the purpose of your assessment and let the decisions that follow support that purpose. Balance the pros and cons, or costs and benefits, of each decision and learn as development proceeds. The field of performance-based assessment is too new for anyone to dictate the right way to do things. Everyone is learning by *doing*.

Traditional approaches to rubric development within the field of writing provide a useful model for rubric design in general. The chart on the next page shows typical steps in the development of a rubric.

The lists for holistic rubric development and analytic rubric development are quite similar. One notable difference is the dependence on real performances for clarification of the score points underlying the holistic scoring continuum. This is an essential way to characterize the score points because examinees may find many different ways to demonstrate achievement at the different levels. Consider, for example, a scoring system that classifies student performance as master or non-master. Essentially, there needs to be only one score point on this continuum. This point represents *master*. An examinee is either at or above that point, or below it.

The point of interest in this scoring scheme is the characterization of the minimal requirement to demonstrate master performance. Because of this relatively gross categorization of performance, it is likely that there will be many, many different ways in which an examinee can demonstrate master performance. In this situation, it may be most efficient for the rubric developers to simply list essential criteria for the master category rather than search for exemplars of the multiple ways to demonstrate this level of achievement.

Determining Score Points

The decision about how many score points are appropriate for a particular assessment activity depends, again, upon the purpose of the assessment and the information needs of the assessment consumers. If master/non-master is the only designation required, one score point is adequate to distinguish between these two

RUBRIC DEVELOPMENT

HOLISTIC	ANALYTIC
Assemble content and grade-level experts	Assemble content and grade-level experts
Review purpose of the assessment	Review purpose of the assessment
Examine samples of student responses to the assessment task	Discuss the number of score points required
Discuss the number of score points required	Specify the elements of the performance to be evaluated
Classify the sample performances into the designated score points or construct prototypes of the different score points	Discuss the characteristics that determine score points for each element
Discuss the characteristics that separate performances into these score points	Identify real responses or write prototypes of each of the different score points
Select or construct exemplar performances for use in training other raters to use the scoring rubric reliably	Select or construct exemplar performances for use in training other raters to use the scoring rubric reliably
Try out the scoring rubric	Try out the scoring rubric
Resolve discrepancies and revise the rubric or exemplars as required	Resolve discrepancies and revise the rubric or exemplars as required
Develop and try out a training program with a focus on level of interrater reliability obtained with given levels of training.	Develop and try out a training program with a focus on level of interrater reliability obtained with given levels of training.

categories of performance. If, however, there is interest in making finer discriminations among examinees for other reasons (including placement, grading), more score points are required.

The number of score points that can legitimately be supported by an assessment depends upon the ability of the assessment developer to define performance standards. That is, each score point must convey with meaning a clearly articulated and differentiated level of performance. As Wiggins (1990) states:

> Standards are specific and guiding pictures of worthy goals. Standards are not abstract aims, wishful thinking, or arcane psychometric tricks. (p. 20)

In reviewing "The Soda Task" scoring guide (Appendix D), note that there are four score points identified: Excellent, Good, Fair, and Poor. In scoring performances, the raters would be trained to implement this scale by examining prototypes or exemplars of each of these score points and, most probably, some performances which seem to fall somewhere in between these score points. These prototypes become a very effective vehicle for communicating precisely to teachers, students, and parents just what these performance standards are.

Voltmer (personal communication, February 1991) suggests that a rubric should have an even number of score points, perhaps four (4) or six (6). She particularly recommends against having five (5) points on the scale because that point system resembles the traditional A, B, C, D, and F grading scale. Having an even number of score points encourages the raters to make discriminating decisions about each performance. With an odd number of score points, the raters may tend to use the midpoint "comfort zone."

Relative to what labels to give the selected score points, there is infinite flexibility. For example, the score points may be described by numbers, letters, words, or phrases. It is important, however, that these labels convey something meaningful to the information users (i.e., teachers, students, parents, administrators). Thus, for example, the score point representing the highest level of achievement might be labeled *Exceptional Achievement* and the lowest level of achievement might be labeled *Inappropriate Response*, following guidelines in the direct writing assessment. Labels such as *Awesome* or *Totally Bad*, while part of the language of this generation of students, may be ambiguous to other users and do not communicate in terms related to the assessment task.

Because performance-based assessment is innovative and likely to seem quite unusual to many constituencies of schools, it seems sensible to communicate performance outcomes in ways that contribute to understanding the value of innovative assessment. For this reason, it is particularly important for developers and users of performance-based assessment to think about both the information captured by the scoring rubric and the information conveyed by the score points. Labels for those score points are primary vehicles for communication.

Articulating Content and Performance Standards

The movement from assessment design to scoring to reporting of performance standards completes a cycle that begins with content standards (See Figure 2).

It is important to remember that there is a difference between content standards and performance standards. Determination of content standards must be made before the stage is set for developing the performance-based assessment. Determination of performance standards is made in the course of developing the scoring rubric.

An example of movement around this cycle comes from the National Assessment of Educational Progress (NAEP). The content standards for the 1990 Mathematics Assessment were derived from the NCTM *Standards*. Though only a few of the test questions used by NAEP are released, the scoring descriptors illustrate the difference between content standards and performance standards.

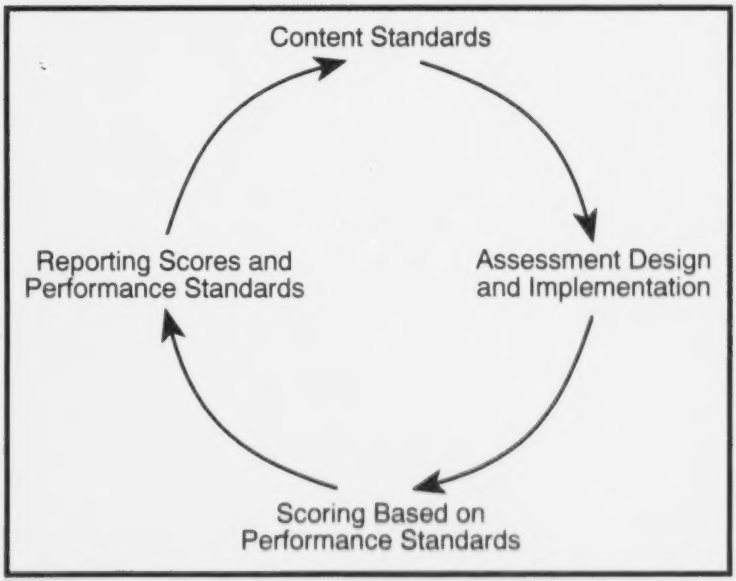

Content Standards
Assessment Design and Implementation
Scoring Based on Performance Standards
Reporting Scores and Performance Standards

Figure 2. Movement from assessment design from content standards to reporting of performance standards.

In *The LEVELS of Mathematics Achievement* (Bourque & Garrison, 1991), three achievement levels are defined: 1) Basic, 2) Proficient, and 3) Advanced. Basic is described as "partial mastery of knowledge and skills." Proficient is described as "solid academic performance." Advanced is described as "superior performance."

In order for consumers to understand these score labels, however, further articulation is necessary. Basic is further defined as denoting partial mastery of knowledge and skills that are fundamental for proficient work. Proficient is defined as representing solid academic performance. Advanced is defined as signifying superior performance beyond proficient grade-level mastery (p. 5).

Even these enhanced descriptions offer very little to the information user. So NAEP has enlarged and elaborated descriptions of the levels in a way that clearly connects the performance standards to the content standards underpinning the assessment:

Basic: Partial Mastery of Knowledge and Skills

Fourth-grade students who are performing at the basic level should be able to solve routine one-step problems involving whole numbers with and without the use of a calculator. They should also be able to use physical materials and pictures to help them understand and explain mathematical concepts and procedures.

Students at this level are beginning to develop estimation skills in measurement and number situations and should understand the meaning of whole number operations. For example, students performing at the basic level should be able to link the meaning of multiplication with the symbols needed to represent it. These students are also beginning to develop concepts related to fractions and read simple measurement instruments. Fourth-grade students performing at the basic level should also be able to identify simple geometric figures and extend simple patterns involving geometric figures. These students should be able to read and use information from simple bar graphs.

Proficient: Solid Academic Performance

Fourth-grade students who are performing at the proficient level should have an understanding of numbers and their application to situations from students' daily lives. The proficient student should be able to solve a wide variety of mathematical problems: use patterns and relationships to analyze mathematical situations; relate physical materials, pictures, and diagrams to mathematical ideas; and find and use relevant information in problem solving. Fourth-grade proficient students should understand the numbers and concepts of place value and have an understanding of whole number operations, as well as a facility with whole number computation. For example, students should be able to solve problems with a calculator and use estimation skills to solve problems. Proficient fourth-grade students should understand and use measurement concepts such as length; be able to collect, interpret, and display data; and use simple measurement instruments.

Advanced: Superior Performance

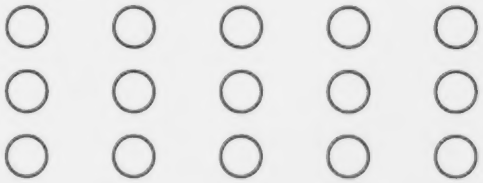
Fourth-grade students who are performing at the advanced level should be able to demonstrate flexibility in solving problems and relating knowledge to new situations. They should be able to use whole numbers to analyze more complex problems. Their understanding of fractions and decimals should extend to a number of representations. Students at this level should determine when estimation or calculator use is an appropriate solution to a problem, as well as read and interpret complex graphs. Advanced fourth-grade students should also be able to use measuring instruments in non-routine ways. These students should be able to solve simple problems involving geometric concepts and chance.

NAEP is clear in differentiating articulation between content standards and performance standards. Furthermore, NAEP provides concrete and explicit examples of precisely how the expected behaviors are translated into test questions (see Examples 1, 2, & 3; note that these questions are not performance-based assessment activities, but they illustrate clarity in communicating behaviors.)

The challenge for developers of performance-based assessments is to have a clear understanding of what constitutes hard content in science and mathematics, and how content standards can be translated into reasonable and appropriate performance standards. Wiggins (1991) suggests, "Real (performance) standards

enable all performers to understand their daily work in terms of specific exemplars for the work in progress, and thus how to monitor and raise their standards" (p.20). The articulation of content standards into real, understandable performance standards is a critical step in the effective use of performance-based assessment in the classroom. It is just as critical that these performance standards be debated openly and communicated widely.

Example #1. Fourth grade, basic level



Write a multiplication sentence to find the number of circles.

3 x 5 = 15

Example #2. Fourth grade, proficient level

On a flight from Los Angeles to New York, the cost of a fare was $400. Every seat was sold. What additional information do you need to find the total for all fares?

A None
B The number of employees on the plane
(C) The number of passenger seats on the plane
D The distance from Los Angeles to New York

Example #3. Fourth grade, advanced level

The table below shows some number pairs. The following rule was used to find each number in column B.

Rule: Multiply the number in column A by itself and then add 3. Fill in the missing number, using the same rule.

A	B
2	7 = (2 x 2) + 3
3	12
5	28
8	67

Chapter 6
How Can Teachers Be Informed Consumers?

This chapter provides guidelines for consumers of performance-based assessments. The guidelines aid in reviewing commercially developed assessments, as well as those developed by colleagues or available in the public domain.

The field of performance-based assessment is growing rapidly. Teachers are beginning to develop performance-based assessments in the classroom, and test publishing companies are marketing assessments that carry the label of *authentic*, *performance*, *performance-based*, or *portfolio assessment*. However, one of the directions in which test developers have not moved very far is toward systematic review procedures for these increasingly complex assessments.

In traditional multiple-choice test development, there are widely accepted and relatively uniform procedures that govern the production of tests. Even in the area of informal or classroom assessment, there are straightforward steps to follow in order to increase the likelihood that the product developed will be reliable, valid, and useful (e.g., Bloom et al., 1971; Popham, 1988). In the area of innovative assessment, however, guidelines are not readily available.

As performance-based assessment continues to grow in acceptance, value, and use, it becomes increasingly important that consumers have some reasonable frameworks for comparing or reviewing these innovative assessments. Of course, frameworks will change as the field becomes more and more sophisticated, but it is helpful for teachers, administrators, critics, publishers, and test professionals to have some initial frames of reference for evaluating performance-based assessments.

One of the reasons that measurement professionals are able to articulate such thorough review procedures for traditional multiple-choice test questions is that there is substantial breadth and depth of experience in the field. That experience is only just beginning to accumulate in the area of performance-based assessment, but as work continues, the tools for review are beginning to emerge.

One reasonable frame of reference, the *Assessment Rating Form*, was presented in Chapter 4. This list of essential elements was used to evaluate "The Soda Task" and provides a sound basis for discussion and reflection.

Another tool, which is a direct outgrowth of work with elementary and middle school teachers designing performance-based assessments in science and mathematics, is the *Performance-Based Assessment Checklist*. As with any frame of reference tool, the *Checklist* has no right or wrong answers; it is intended to guide consumers through the review process in a systematic manner. The only right answers are those that are right for the specific context and purpose for which the performance-based assessment is being reviewed.

If one were evaluating "The Soda Task" for possible use in a eleventh-grade chemistry class, the *Performance-Based Assessment Checklist* might be completed as follows:

PERFORMANCE–BASED ASSESSMENT CHECKLIST[1]

Name of Task: "The Boat Task"

Name of School: Anytown High School

Grade Level: 11 **Date:** 5/94 **Task Code:**

Subject Area(s): Chemistry

Evaluator: Margaret Jamison

1. *What is the topic to be focused on?* Scientific investigation

2. *Is the topic broad enough that assessments beyond knowledge of factual material can be developed?* Yes

3. *What are the goals for the Performance-Based Assessment Task?*
Students should be able to meet the 11 objectives listed in Chapter 4.

4. *Check the levels in each of the domains that the Performance-Based Assessment Task addresses.*

PSYCHOMOTOR[1]	AFFECTIVE[2]	COGNITIVE[3]
Perception	Receiving ✔	Knowledge ✔
Set	Responding ✔	Comprehension ✔
Guided Response	Valuing	Application ✔
Mechanism	Organization	Analysis ✔
Complex Overt Response	Value Complex	Synthesis ✔
Adaptation		Evaluation ✔
Origination		

5. *Check the skills that the Performance-Based Assessment Task allows the student to demonstrate.*

Classifying ✔		
Communicating:	Speaking ✔	Listening ✔
	Reading ✔	Writing ✔
Constructing Hypotheses ✔		
Cooperation/Collaboration ✔		
Creative Thinking ✔		
Critical Thinking ✔		
Data/Information:	Locating ✔	Organizing ✔
	Analyzing/Interpreting ✔	Evaluating ✔
Defining Operationally ✔		
Drawing Conclusions ✔		

Experimenting/Investigating ✔
Formulating Models ✔
Identifying and Manipulating Variables ✔
Inferring ✔
Interpreting Literature NA
Measuring ✔
Observing ✔
Predicting ✔
Problem Solving: Identifying Problems ✔ Formulating Problems ✔
Formulating Possible Solutions ✔
Choosing Optimal Solutions ✔
Evaluating Results ✔
Complex Problems ✔ Multistep Problems ✔
Synthesizing Knowledge From a Variety of Sources ✔
Using Mental Computation Strategies ✔
Using Estimation Strategies ✔
Using Map and Globe Skills NA
Using Reference and Study Skills ✔
Using Space/Time Relationships NA

6. *Does the Performance-Based Assessment Task call for:*

active student learning?	Yes ✔	No
divergent thinking?	Yes ✔	No
holistic activities?	Yes ✔	No

7. *Is the Performance-Based Assessment Task:*

• at the appropriate level of difficulty?	Yes ✔	No
• feasible for implementation within the constraints under which the teacher must work (space and equipment, time, and types of students)?	Yes ✔	No
• feasible for the student to complete the activity with a sense of closure and accomplishment?	Yes ✔	No
• cost effective?	Yes ✔	No
• guided by clear directions to the teacher?	Yes ✔	No
• guided by clear directions to the student?	Yes ✔	No

8. *Does the Performance-Based Assessment Task have:*

• multiple goals?	Yes ✔	No
• activities that allow for integration across different subject areas?	Yes ✔	No
• motivational value?	Yes ✔	No
• activities that are constructed around currently or recently taught powerful ideas at an appropriate place in the curriculum?	Yes ✔	No

- activities that challenge the student not just to locate and reproduce information but to interpret, analyze, or manipulate information in response to a question or problem that cannot be resolved through routine application of previously learned knowledge? Yes ✔ No
- activities that can be adapted to accommodate individual differences in interests or abilities? Yes ✔ No
- variety? Yes ✔ No
- progressive levels of difficulty or complexity? Yes ✔ No
- life applications? Yes ✔ No
- full range of goals addressed? Yes ✔ No
- concrete experiences? Yes ✔ No
- activities that connect declarative knowledge with procedural knowledge? Yes ✔ No
- valid content? Yes ✔ No
- extension activities? Yes ✔ No

9. *Has a rubric (scoring guide) been designed?* Yes ✔ No

   If **Yes**, is it holistic, analytic, 4 or mastery?
   How many points? 4
   How many scores? 9

10. *Is the rubric consistent with the stated purpose?* Yes ✔ No

11. *Are data available?* Yes ✔ No

    If **Yes**, where?
    Have results been reported? Yes ✔ No
    in what form? with what implications?[5]

12. *Have standards been set?* Yes No ✔
    If **Yes**, what are they?

13. *Will the results be meaningful to the users?* Yes ✔ No
    If **Yes**, how?

14. *Initial Small-Scale Tryout Report:*

15. *Field-Test Report:*

---

[1] *Original conception and design by Ellen Marie Moore, Independent Consultant to ETS, Rising Fawn, GA, 1991. ETS expresses gratitude for this important work.* *(Copyright - 1991 by Educational Testing Service. All rights reserved.)*

[2] (Simpson, 1966)

[3] (Krathwohl et al., 1964)

[4] (Bloom et al., 1956)

[5] (Baron, 1990, 1991; Baron et al., 1989)

There is no "key" to this questionnaire because the *rightness* or *wrongness* of the responses is a function of the purpose for which the performance-based assessment is being reviewed. In the case of assessments developed without the support of a major research project, the match between the assessment characteristics and the *Performance-Based Assessment Checklist* is not likely to be as complete as it is with "The Soda Task." However, the process of going through these questions can be of value in refining the assessment and identifying gaps in the development process, as well as in providing a basis for adoption decisions.

This *Checklist* is not intended to be a scoring rubric. Do not add up the total number of Yes responses, for example, to give the assessment under review a score. These questions are not of equal value. They serve as a frame of reference for decision making. They may identify an area that was not covered, prompt the expansion of the assessment in certain ways, or confirm the direction undertaken. The overriding question, of course, is whether or not the assessment meets the needs of the user.

Linn et al. (1991) provide still another frame of reference for reviewing performance-based assessments. They address some different aspects of the assessment activity that may be useful in combination with either the *Assessment Rating Form* (see Chapter 4) or the *Checklist* above. These authors suggest the following criteria for reviewing performance-based assessment activities:

- Consequences of Use
- Fairness
- Transfer and Generalizability
- Cognitive Complexity
- Content Quality
- Content Coverage
- Meaningfulness
- Cost and Efficiency

Some of these characteristics are consistent with areas covered in the *Performance-Based Assessment Checklist*. Some are not. For example, "Consequences" and "Fairness" are not included in the *Checklist* but are important criteria to consider.

A review of "The Soda Task," for instance, should thoroughly examine the *consequences* of implementing this type of assessment. This question might be answered in terms of instruction: Would this assessment serve as a powerful model for high-quality instruction, or does it support rote memory or routine problem solving? Does this assessment have positive consequences for learning? What are the consequences for use of the score information? Would scores based upon collaborative work be interpreted as individual data, or would information about individual performance be interpreted as problem solving when it represents, instead, recitation of others' ideas?

In terms of *fairness*, does "The Soda Task" represent the kind of instructional activity that the majority, if not all, of the students have had during their chemistry

course(s)? "The Soda Task" cannot be a fair and equitable assessment for all students if students have had unequal access to practice in this type of activity.

In terms of *transfer and generalizability*, opportunity is provided to see whether skills demonstrated on "The Soda Task" are also evident in similar or parallel activities. In terms of *cognitive complexity*, there is little doubt that the task taps higher-order thinking skills. In terms of *content quality, coverage,* and *meaningfulness,* "The Soda Task" engages students in the full range of processes and skills that underlie what scientists do.

Finally, in terms of *cost and efficiency*, it seems reasonable to assume that the equipment required for "The Soda Task" is readily available in high school chemistry classrooms, and any special materials are inexpensive. Whether this assessment activity is the most efficient for the purpose is hard to tell. It is possible that other, less expensive, and shorter assessment activities would yield comparable data about individuals and groups. Surely there will come a time when these types of assessments will be conducted through simulation or virtual reality at a minimal cost of time, dollars, and facilities. Given the state of the art in performance-based measurement at this time, however, "The Soda Task" seems to be quite reasonable in terms of both cost and efficiency.

These three frames of reference are offered as alternative and somewhat complementary tools for reviewing performance-based assessments whether they be locally developed or commercially available. The bottom line, regardless of which review strategy or combination of strategies is selected, is that the assessment must satisfy the purposes for which it is being used, and the information disseminated relative to performance on the assessment must be accurate and meaningful.

Chapter 7
What Are Critical Questions About Performance-Based Assessment?

This chapter closes the book with a presentation of important but, as yet, unanswered questions. These include questions of equity, fairness, consequences of use, and concerns shared by theoreticians and practitioners about the value and appropriateness of performance-based assessment.

As Ruth Mitchell, a well-recognized contributor to the field of innovative assessment, says: "Alternative assessments can take as many forms as imagination will allow" (cited in Willis, 1990, p. 4). Mitchell's statement is a powerful reminder that innovative assessment requires creativity. It requires a paradigm shift in how one thinks about tests. It requires a paradigm shift in terms of how tests function within the culture of the school. It requires a paradigm shift in the role of teachers in the relationship between instruction and assessment.

Making these paradigm shifts while being creative is one of the challenges in the area of performance-based assessment. It takes time to be creative, and time is money. But as Mitchell reminds us, the limits for this type of assessment are defined only by one's imagination. It is exciting to have the opportunity, the flexibility, and the challenge for creativity both in the professional measurement community and in the classroom.

It is critical to the emerging field of performance-based assessment that classroom teachers remain involved as developers, scorers, and critics. Without the classroom as a research site for the development and refinement of performance-based assessment, there are likely to be fundamental flaws in the assessments. The act of juggling traditional job responsibilities and the additional challenges of developing performance-based assessments will also require a paradigm shift. Because it is imperative that teachers and administrators remain actively involved in assessment development, some relief from responsibilities in other areas must be sought. Involvement in performance-based assessment must not become "just one more thing for busy teachers to add to their day."

In becoming partners with the professional measurement community, educators (teachers and administrators) must shift from a practitioner paradigm to a research paradigm. There are many more unanswered questions about performance-based assessment than there are answered ones. Classroom teachers, in particular, are the key to answering these questions because of their insights and daily experience with instruction and assessment.

Some of the issues on the research agenda for performance-based assessment are:

- What is the difference between instruction and assessment?
- Will performance-based assessment enable students who know the content but perform poorly on traditional tests to demonstrate more clearly what they know?
- Will performance-based assessment avoid the pitfalls of traditional tests?
- What is the impact of student collaboration on scores?
- What are the cost implications of performance-based assessment and what are the payoffs?
- Can scoring be accomplished efficiently and accurately?
- What are the reporting constraints/requirements for performance-based assessment?
- What will ensure the connection between instruction and performance-based assessment?
- How manageable is performance-based assessment?
- What are the implications for teacher training inherent in the use of performance-based assessment?
- Will a new generation of psychometric theory be developed to accompany performance-based assessment?

The discussion that follows elaborates on this agenda.

◆ *What is the difference between instruction and assessment?*

There is a fundamental tension between using performance-based assessment in instruction and using it to evaluate accountability. Frequently, in fact, it is espoused for both. One cannot read *America 2000* or listen to reports from the National Education Goals Panel without sensing the push to use performance-based assessment for both purposes.

Regardless of whether a test is multiple choice or performance based, the two purposes of assessment remain distinct. Simply applying the principles of matrix sampling will not convert an instructional assessment into an accountability assessment. There is not likely to be one approach or methodology that will meet the needs of both adequately.

Historically, accountability assessment is a top-down testing program that holds states, school systems, schools, and sometimes teachers accountable for specific levels of learning. The 1970s and 1980s are full of examples of statewide criterion-referenced or competency-based tests used for promotion, retention, and credentials for high school completion (i.e., certificate of attendance versus diploma).

These tests clearly determined where teachers placed their instructional emphasis, because the stakes for students, as well as educators, were high. So, teachers spent weeks preparing their students for the tests, with drill and practice in test taking as well as in the content to be covered. As Madaus et al. (1992) state, this concern about top-down accountability measures seriously conflicted with what good teachers wanted to do in science and mathematics classrooms.

Although this text has used the phrase *performance-based assessment* as an umbrella term to include many other descriptors, Meyer's (1992) definition of authentic assessment provides an interesting perspective on instruction and assessment. She states, "In an authentic assessment, the student not only completes or demonstrates the desired behavior, but also does it is a real-life context" (p. 40).

From Meyer's perspective, authenticity may be a useful element when describing the difference between instructional and accountability assessments. After all, can an assessment really be authentic for a diverse student population and still yield data that can be aggregated? Moreover, doesn't traditional accountability assessment assume aggregation at the classroom, school, and school system level?

Another way to look at the distinction is to ask, who is in control? From Meyer's perspective, control must reside with the student if the assessment is to be authentic. That certainly works in an instructional assessment setting. But what about in an accountability setting? Certainly, in the latter context, the test administrator must be in control.

There is now beginning to emerge, however, another understanding of accountability, that is, the accountability of the student for managing his or her own learning. Part of that management paradigm relies on assessment. Thus, as the paradigm shifts from top-down to bottom-up accountability assessment, the distinction between instruction and assessment becomes quite blurred.

These are interesting issues that will continue to be debated vigorously over the next few years. Researchers and practitioners are only beginning to understand what limits, if any, are necessary on performance-based instructional assessment to ensure validity and reliability. Clarifying the fuzzy distinction between instruction and assessment within that context has only just begun.

- *Will performance-based assessment enable students who know the content but perform poorly on traditional tests to demonstrate more clearly what they know?*

There is a strong belief that the equity issues inherent in traditional testing will disappear as testing moves towards performance-based assessment. Furthermore, "true achievement" will manifest itself in performance-based environments, and students who really know and understand but who cannot respond correctly in a structured multiple-choice environment will flourish. The extent to which performance-based assessments facilitate or inhibit the demonstration of learning must be researched thoroughly before one practice is abandoned in favor of another. It would be a serious disservice to students if the move from traditional assessment to innovative assessment were based on belief rather than on research and if the actual impact resulted in yet more inequities.

Research is beginning to suggest that generalizability about an individual student's achievement in a defined content domain must be based on 10–20 different performance-based assessment activities (Shavelson & Baxter, 1992). In essence, a single performance-based assessment activity can be considered equivalent to a rather short multiple-choice test. It is well-known in classic measurement

theory that the longer the test sample is, the more reliable the results are. This is intuitively logical. Longer exposure to how a student performs will always provide more stable information than brief snapshots.

A question that remains for researchers in performance-based assessment is whether or not systematic observation of student behavior summarized by a numerical or category description (i.e., a performance-based assessment with scoring rubric), in combination with performance in the instructional setting or other assessment samples, will reduce the required activities to a less cumbersome number. If not, it is unlikely that there will be the resources necessary to assess each valued outcome with the 10–20 different performance-based activities necessary to produce reliable information.

- *Will performance-based assessment avoid the pitfalls of traditional tests?*

Of the many so-called pitfalls attributed to multiple-choice testing, principal ones include content coverage, equity, and trickiness. These pitfalls are as likely to be present in poorly designed performance-based assessments as in poorly designed multiple-choice tests. In short, it is not the tool that has inherent pitfalls; it is the weakness in human design and thought.

In terms of content coverage, the fact that performance-based assessments tap multiple goals does not mean that they do not *sample* the curriculum in much the same way that multiple-choice tests do. In fact, short of nonstop testing, a situation in which assessment becomes instruction, there seems no way to avoid assessment that samples instruction. How then can assurance be obtained that performance-based assessment captures deeper understanding, more effective transfer, and higher-order thinking?

In terms of equity, if the assessment captures evidence about behaviors never practiced and content never taught or learned, equity problems will persist, regardless of the format of assessment. It is, however, particularly critical that educators and measurement professionals involved in performance-based assessment make certain that assessments are used in conjunction with high-quality instruction that provides practice in performance tasks and promotes higher-order thinking skills. To the extent that performance-based assessment is used in classrooms where traditional instruction prevails, the question of equity will have validity. One must then ask how performance-based assessment and instructional reform can be introduced in tandem and paced to complement each other?

- *What is the impact of student collaboration on scores?*

When performance-based assessments are used to generate a score, how will the setting, the grouping, and the extent of teacher intervention/participation be factored in? Will scores given to collaborative assessments be somehow weighted by other variables? If group scores are assigned to individuals, will individuals have the freedom to select the members of their group? If some groups require more assistance from the teacher than others, will this assistance be somehow quantified or qualified and used to reduce the group achievement estimate?

◆ *What are the cost implications of performance-based assessment and what are the payoffs?*

Preliminary information reported by R. Hill (personal communication, April 1992) from his experience as project director for the assessment component of the Kentucky Education Reform Act is that it costs about 10 times as much to develop a performance-based assessment activity as to develop a typical multiple-choice test. One must add to this development cost the cost of assembly, packaging, shipping, scoring, and reporting.

If Hill's estimate of cost is accurate and generalizable from one testing company to another, and if cost reductions do not result from increased experience, the financial burden of developing performance-based assessment for use in statewide, relatively high-stakes testing programs will be prohibitive for most parts of the country.

How this cost estimate translates into cost for teacher design and classroom use for informing instruction is unknown because of the absence of systematically collected data. Based upon this author's experience in training teachers and administrators to design performance-based assessments for classroom use, it is reasonable to expect that it takes a minimum of three days of staff development and hands-on assessment design to construct a reasonable draft of the assessment activity and rubric for small-scale tryout.

When a period of three days is compared to the amount of time teachers typically spend on writing informal teacher-made tests, the real cost of performance-based tests becomes profoundly apparent. Clearly, most teachers who use performance-based assessments will have to purchase these assessments rather than develop them.

In terms of payoffs, many would take the philosophical position that performance-based assessments must be used often in the classroom because of what they symbolize to teachers and students and because of what they model. If this is the position taken, every effort must be made to ensure that the quality of performance-based assessments, whether bought, borrowed, adapted, or developed, sends clear and positive signals about higher-order thinking and the products of schooling.

◆ *Can scoring be accomplished efficiently and accurately?*

From the history of the direct assessment of writing at statewide levels, it is clear that people can be trained to score writing samples reliably. It is also clear that enough raters are available for even massive testing programs that require human judgment. However, as performance-based assessments extend into relatively content-dependent areas such as trigonometry, calculus, synthetic geometry, physics, and chemistry, the question arises as to the availability of sufficient numbers of appropriately qualified individuals to serve as scorers.

If performance-based assessments are used to measure achievement across classrooms, schools, systems, and states, can the scoring be done feasibly in terms

of time, money, or expertise? Where will the experts be found to evaluate performance-based assessments that are complex, interdisciplinary activities? Can the scoring be completed in a time frame short enough so that the immediate benefits will be felt in the classroom? Or will the need for objective, controlled scoring typically required in accountability assessments override the needs of teachers and students?

- *What are the reporting constraints/requirements for performance-based assessment?*

In a review of scoring and reporting rubrics used in *Looking Beyond the Answer* (Vermont Department of Education, 1991) and *A Question of Thinking* (California Department of Education, 1989), it is interesting to note that the rubrics in and of themselves tend not to be descriptive of content standards. However, the exemplars provided for each score point are descriptive of both content and performance standards. This reinforces the notion that generalizable scoring rubrics, if supported by content-specific exemplars of score points, should effectively reduce the amount of development time, make uniform the reporting framework, and generally expedite the scoring and reporting process. Whether or not single rubrics could meet the needs of multiple information users is a question to be researched.

- *What will ensure the connection between instruction and performance-based assessment?*

What will tie performance-based assessment to instruction? Will teachers value the assessments enough to integrate them into their teaching programs? Will the assessments provide such enlightening models that teachers and students will internalize their characteristics through exposure? Or will teacher training, either preservice or in-service, be necessary to incorporate performance-based assessment effectively into classroom instruction? If this latter be the case, both teacher training programs and staff development programs must move quickly to prepare teachers for the new assessments.

- *How manageable is performance-based assessment?*

In terms of implementing or administering the assessments, there is no evidence that management is an issue. Even with young children (grades K-3) when manipulatives are involved, both teachers and students report handling the situation easily (Hardy, 1992).

Shipping and distribution are more complicated than in traditional multiple-choice testing simply because there is typically more to distribute than booklets and answer sheets. Beyond the relative bulk of the materials and the associated costs, however, the management of the distribution seems straightforward.

With regard to scoring the responses, assembling and training human beings to make reliable and accurate judgments about performance will always be more

problematic than running sheets through a scanner. Moreover, the potential scarcity of expertise in some of the content areas may make scoring sessions difficult to plan. However, for performance-based assessments used in the classroom, where the teacher or a local colleague is likely to be the rater, the scarcity issue is not likely to arise.

On another dimension related to scoring, will it be possible to create transferable scoring rubrics which can be applied across different performance-based assessment tasks? Does each distinct performance-based assessment require a customized scoring rubric or can generic rubrics be developed which will provide sufficient detail? The implications for manageability of the scoring process and assimilation of the information derived are substantial.

- *What are the implications for teacher training inherent in the use of performance-based assessment?*

Teacher training institutions are currently taking a slow and cautious approach to infusing performance-based assessment theory and practice into teacher-training programs. Currently, the majority of training provided in this area is at the school or system level through in-service or staff development programs.

If novice teachers are to adopt performance-based assessment at the classroom level, they must be provided theoretical and practical experience or, at the very least, exposure to fundamental principles of measurement and the emerging literature on performance-based assessment. This investment in young teachers could yield an impressive return as these individuals join the profession and make immediate contributions to the use of performance-based assessments.

- *Will a new generation of psychometric theory be developed to accompany performance-based assessment?*

In traditional assessment, the measurement community has developed sophisticated methodology to ensure that certain assessments can be substituted for others with complete fidelity. This knowledge base is missing in performance-based assessment.

Should performance-based assessment be classified as a new measurement field with a need to devise equating strategies so that this same kind of substitution is possible? Is enough known to ascertain that pre- and post-assessments are indeed measuring the same thing? Is enough known to have confidence in growth or trend data generated using performance-based assessment, or would users and critics alike question the comparability?

Indeed, this dilemma may be the critical one in that the development activities required for performance-based assessments have rapidly moved ahead of the psychometric thought in this area. Perhaps it is time to begin to examine the psychometric properties of these innovative assessments to determine if concepts like reliability and validity can be documented in the traditional ways or

whether performance-based assessment requires innovative statistical methodology as well. Test developers would be well served to address this question quickly before policy decisions begin to form on the basis of theoretically idiosyncratic assessments.

In *America 2000*, President Bush said:

> Nothing better defines what we are and what we will become than the education of our children....If we want America to remain a leader, a force for good in the world, we must lead the way in educational innovation....Think about every problem, every challenge we face. The solution to each starts with education....The days of the status quo are over....To those who want to see real improvement in American education, I say: There will be no renaissance without revolution. (pp. 1–3)

It is truly the case that there is revolution in assessment. Performance-based assessment requires revolutionary thought about learning and about assessment, about what and how students should be evaluated, and about who is responsible for learning and who is accountable for learning. As more and more teachers become increasingly active in this revolution, there can be only positive results.

The measurement and education communities are in the throes of that revolution now. Performance-based assessment offers a unique opportunity to move forward, to enhance the process of schooling, and to make a difference in how citizens view education in the United States. But the rhetoric and the intuitive appeal must be supported by research and careful investigation.

As Mitchell says:

> Alternative assessments also serve the goal of greater teacher empowerment by allowing teachers to play a central role in designing, administering, and scoring assessment tasks. These efforts are the world's best form of professional development because they make teachers carefully consider what they want their students to know and how they can ensure that students have learned it. (cited in Willis, 1990, p. 4)

In science and mathematics the revolution was ignited by *Science for All Americans* and the NCTM *Standards*. It is up to practicing teachers and measurement professionals actively involved in science and mathematics education to move beyond rhetoric to sound practice.

References

American Association for the Advancement of Science. (1989). *Science for all Americans*. Washington, DC: Author.

Anrig, G. R. (1991, Spring/Summer Special Edition). President's message: Commitment to improved learning and opportunity. *ACCESS*, pp. 1, 11.

Anrig, G. R. (1992, April). *Being put to the test in education*. Paper presented at the Convention of the National School Boards Association, Orlando, FL.

Baker, E. L. (1990, October). What probably works in alternative assessment. In *The promise and peril of alternative assessment*. Conference sponsored by the U.S. Office of Educational Research and Improvement, Washington, DC.

Barker, J. A. (1992). *Future edge: Discovering the new paradigms of success*. New York: William Morrow.

Baron, J. B. (1990). Performance assessment: Blurring the edges among assessment, curriculum, and instruction. In A. B. Callinger, B. E. Lovitts, & B. J. Callinger (Eds.), *This year in school science: Assessment in the service of instruction* (pp. 127-148). Washington, DC: American Association for the Advancement of Science.

Baron, J. B. (1991). Strategies for the development of effective performance exercises. *Applied Measurement in Education, 4*(4), 312-318.

Baron, J. B., Forgione, P. D., Rindone, D. A., Kruglanski, H., & Davey, B. (1989, April). *Toward a new generation of student outcome measures: Connecticut's Common Core of Learning Assessment*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Blank, R. K., & Dalkilic, M. (1992, May). *State policies on science and mathematics education*. Washington, DC: Council of Chief State School Officers.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals, Handbook 1: Cognitive domain*. New York: David McKay.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.

Bourque, M. L., & Garrison, H. H. (1991). *The LEVELS of mathematics achievement: Initial performance standards for the 1990 NAEP mathematics assessment*. Washington, DC: National Assessment Governing Board.

California Department of Education. (1989). *A question of thinking.* Sacramento: Author.

Chira, S. (1991, March 31). The big test: How to translate talk into school reform. *Tallahassee Democrat/Sun*, p. 1.

Chittenden, E. (1990, April). *Authentic assessment, evaluation, and documentation of student performance.* Paper presented at an invitational symposium sponsored by the Association for Supervision and Curriculum Development, San Jose, CA.

Connecticut Department of Education. (1989). The soda task. In *Connecticut Common Core of Learning, Performance Assessment Project.* Hartford: Author.

Cronbach, L. J. (1970). *Essentials of psychological testing.* New York: Harper & Row.

Educational Testing Service. (1991, October). Why are we pursuing new modes of assessment? *New Mode News 2*(1), 1-2.

Finn, C. E. (1990, April). The biggest reform of all. *Phi Delta Kappan, 71*(8), 584-592.

Goals 2000: Educate America Act of 1993 (S. 1150 & H.R. 1804, 103rd Congress).

Hardy, R. (1992). *Options for scoring performance assessment tasks.* San Francisco: National Council on Measurement in Education.

Herman, J. L. (1992, May). What research tells us about good assessment. *Educational Leadership, 49*(8), 74-78.

Hogan, T. P., & Mishler, C. (1982). *Relationships among measures of writing skill.* Washington, DC: National Institute of Education.

Jorgensen, M. (1991). Reflections of learning* (Staff development program). Princeton, NJ: Educational Testing Service.

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives: The classification of educational goals, Handbook 2: Affective domain.* New York: David McKay.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Lindquist, E. F. (1951). Preliminary considerations in objective test construction. In E. F. Lindquist (Ed.), *Educational measurement.* Washington, DC: American Council on Education.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991, November). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Madaus, G. F., West, M. M., Harmon, M., Lomax, R., Viator, K., Fongkong Mungal, C., Butler, P., McDowell, C., Simmons, R., & Sweeney, E. (1992). *The influence of testing on teaching math and science in grades 4-12* (NSF Grant No. SPA8954759). Boston: Boston College, Center for the Study of Testing, Evaluation, and Educational Policy.

McCartin, K. (1992, October 29). Virtual reality: Lobbying for change in America's education system. *Trenton Times*, p. E1.

Meyer, C. A. (1992, May). What's the difference between "authentic" and "performance" assessment? *Educational Leadership, 49*(8), 39-40.

Mitchell, R. (1989, December). *What is performance assessment?* Washington, DC: Council for Basic Education.

Mullis, I.V.S., Dossey, J. A., Foertsch, M. A., Jones, L. R., Gentile, C. A. (1991, November). *Trends in academic progress: Achievement of U.S. students in science, 1969-70 to 1990; mathematics, 1973 to 1990; reading, 1971 to 1990; writing, 1984 to 1990.* Washington, DC: Educational Testing Service/National Center for Education Statistics.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform.* Washington, DC: U.S. Department of Education.

National Council of Teachers of Mathematics. (1980). *An agenda for action.* Reston, VA: Author.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author

National Council of Teachers of Mathematics. (1991). *Professional standards for teaching mathematics.* Reston, VA: Author.

National Research Council. (1989). *Everybody counts: A report to the nation on the future of mathematics education.* Washington, DC: National Academy Press.

Popham, W. J. (1988). *Educational evaluation* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Porter, A. C., Archbald, D. A., & Tyree, A. K. (1991). Reforming the curriculum: Will empowerment policies replace control? In S. H. Fuhrman & B. Malen (Eds.), *The politics of curriculum and testing* (pp. 11-36). London: Falmer Press.

Resnick, L. B. (1987). *Education and learning to think.* Washington, DC: National Academy Press.

Resnick, L. B., & Resnick, D. P. (1989). Tests as standards of achievement in schools. In *Proceedings of the 1989 ETS Invitational Conference* (pp. 63-80). Princeton, NJ: Educational Testing Servce.

Rutherford, F. J., & Ahlgren, A. (1990). *Science for all Americans.* New York: Oxford University Press.

Salinger, T. (1992). Information approaches to assessment. In C. Hedley, D. Feldman, & P. Antonacci (Eds.), *Literacy across the curriculum.* Norwood, NJ: Ablex.

Shavelson, R. J., & Baxter, G. P. (1992, May). What we've learned about assessing hands-on science. *Educational Leadership, 49*(8), 20-23.

Shepard, L. A. (1991, October). Psychometricians' beliefs about learning. *Educational Researcher, 20*(7), 2-16.

Simpson, E. J. (1966). The classification of educational objectives, psychomotor domain. *Illinois Teacher of Home Economics, 10,* 110-144.

Slavin, E. R. (1991). *Student team learning: A practical guide to cooperative learning* (3rd ed.). Washington, DC: National Education Association.

Stalnaker, J. M. (1951). The essay type of examination. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 495-530). Washington, DC: American Council on Education.

U.S. Department of Education. (1991, April). *America 2000: An education strategy.* Washington, DC: U.S. Government Printing Office.

Vermont Department of Education. (1991). *Looking beyond the answer.* Montpelier: Author.

Wiggins, G. (1990, December). *The case for authentic assessment.* Washington, DC: ERIC Clearinghouse on Tests, Measurement, and Evaluation. (ERIC Document Reproduction Service No. ED 328 611)

Wiggins, G. (1991, February). Standards, not standardization: Evoking quality student work. *Educational Leadership, 48*(5), 20.

Wiggins, G., Browne, J., & Houston, H. (1991). *Standards, not standardization.* Stow, MA: Greater Insight Productions.

Willis, S. (1990, September). Transforming the test. *ASCD Update, 32*(7), 3-6.

Zessoules, R., & Gardner, H. (1991). Authentic assessment: Beyond the buzzword and into the classroom. In V. Perone (Ed.), *Expanding student assessments* (pp. 47-71). Alexandria, VA: Association for Supervision and Curriculum Development.

Appendix A

STATE MATHEMATICS CURRICULUM FRAMEWORK OR GUIDE

State	State Framework or Guide Revised with NCTM Standards/Date of Completion	Framework or Guide Relationship to Mathematics Student Assessment	Framework or Guide Relationship to Mathematics Texts
ALABAMA	Yes 1989	DIRECT	RECOMMEND
ALASKA	Revising 1992	INDIRECT	No
ARIZONA	Revising 1992	DIRECT	RECOMMEND
ARKANSAS	Revising 1992	DIRECT	SELECT
CALIFORNIA	Yes 1991	INDIRECT	SELECT
COLORADO	—	LEARNING OUTCOMES	No
CONNECTICUT	Revising 1993	DIRECT	No
DELAWARE	Revising 1992	DIRECT	RECOMMEND
DIST. OF COLUMBIA	Revising 1992	DIRECT	SELECT
FLORIDA	Yes 1991	LEARNING OUTCOMES	SELECT
GEORGIA	Yes 1988	INDIRECT	RECOMMEND
HAWAII	Revising 1992	INDIRECT	RECOMMEND
IDAHO	Yes 1990	Developing Assessment	SELECT
ILLINOIS	Yes 1985 Revising 1994	DIRECT (1994)	No
INDIANA	Yes 1991	INDIRECT	SELECT
IOWA	Revising 1992	Developing new assessment	No
KANSAS	Yes 1990	DIRECT	No
KENTUCKY	Yes 1992	LEARNING OUTCOMES	LEARNING OUTCOMES
LOUISIANA	Revising 1993	DIRECT	RECOMMEND
MAINE	—	—	—
MARYLAND	Yes 1985	DIRECT	No
MASSACHUSETTS	Developing 1994	—	—
MICHIGAN	Yes 1991	DIRECT	No
MINNESOTA	Yes 1991	DIRECT	No
MISSISSIPPI	Revising 1993	DIRECT	SELECT
MISSOURI	Yes 1991	DIRECT	RECOMMEND
MONTANA	Developing 1994	—	—
NEBRASKA	—	—	—
NEVADA	Yes 1992	INDIRECT	RECOMMEND
NEW HAMPSHIRE	—	—	—
NEW JERSEY	Yes 1990	LEARNING OUTCOMES	No
NEW MEXICO	Revising 1992	INDIRECT	No
NEW YORK	Yes 1990	DIRECT	No
NORTH CAROLINA	Revising 1992	DIRECT	SELECT
NORTH DAKOTA	Developing 1992	—	No
OHIO	Yes 1991	DIRECT	No
OKLAHOMA	Yes 1991	DIRECT	SELECT

STATE MATHEMATICS CURRICULUM FRAMEWORK OR GUIDE

State	State Framework or Guide Revised with NCTM Standards/Date of Completion	Framework or Guide Relationship to Mathematics Student Assessment	Framework or Guide Relationship to Mathematics Texts
OREGON	Yes/1987	DIRECT	SELECT
PENNSYLVANIA	—	—	—
RHODE ISLAND	Developing/1993	—	—
SOUTH CAROLINA	Revising/1994	DIRECT	SELECT
SOUTH DAKOTA	—	—	—
TENNESSEE	Yes/1991	DIRECT	SELECT
TEXAS	Yes/1991	DIRECT	RECOMMEND
UTAH	Revising/1992	DIRECT	SELECT
VERMONT	Revising/1993	Developing assessment	—
VIRGINIA	Yes/1988	INDIRECT	SELECT
WASHINGTON	Yes/1991	No response	No Response
WEST VIRGINIA	Yes/1991	LEARNING OUTCOMES	SELECT
WISCONSIN	Revising/1993	INDIRECT	No
WYOMING	Yes/1990	No state assessment	No
TOTAL	Yes/Revising=41 States Developing=4 States No=6 States	DIRECT=22 States INDIRECT=10 States LEARNING OUTCOMES=5 States	SELECT=15 States RECOMMEND=9 States

DIRECT = *Direct linkage between framework or guide and assessment, i.e. curriculum framework defines content topics and skills to be assessed in mathematics.*
INDIRECT = *Curriculum framework defines goals or objectives for instruction, and assessment is developed or selected to reflect goals and objectives.*
LEARNING OUTCOMES = *State has desired learning outcomes, separate from curriculum framework, and the learning outcomes are used to develop the student assessment.*
SELECT = *Mathematics curriculum guide or framework is used to select state-approved textbooks.*
RECOMMEND = *Mathematics curriculum guide or framework is used to recommend a list of textbooks, with selection being made by local districts.*
— *No state curriculum framework or guide.*

Source: State Department of Education, Mathematics and Science Supervisors, Winter, 1992.

Appendix B

STATE SCIENCE CURRICULUM FRAMEWORK OR GUIDE

State	Science Framework or Guide Date of Completion	Framework or Guide Relationship toScience Student Assessment	Framework or Guide Relationship to Science Textbooks
ALABAMA	Yes/1988	DIRECT	RECOMMEND
ALASKA	Revising/1994	No state assessment	No
ARIZONA	Yes/1990	DIRECT	RECOMMEND
ARKANSAS	Yes/1990	DIRECT	SELECT
CALIFORNIA	Yes/1990	INDIRECT	SELECT
COLORADO	—	LEARNING OUTCOMES	No
CONNECTICUT	Yes/1991	DIRECT	No
DELAWARE	Developing/1994	No state assessment	RECOMMEND
DIST. OF COLUMBIA	Developing/1993	—	SELECT
FLORIDA	Yes/1990	No state assessment	SELECT
GEORGIA	Yes/1988	INDIRECT	RECOMMEND
HAWAII	Revising/1992	INDIRECT	RECOMMEND
IDAHO	Yes/1989	No state assessment	SELECT
ILLINOIS	Yes/1985,Revising/1994	DIRECT (1994)	No
INDIANA	Developing/1992	DIRECT	SELECT
IOWA	Yes/1991	Developing new assessment	No
KANSAS	Developing/1993	Developing learning outcomes	No
KENTUCKY	Developing/1993	LEARNING OUTCOMES	LEARNING OUTCOMES
LOUISIANA	Yes/1991	DIRECT	RECOMMEND
MAINE	—	—	—
MARYLAND	Yes/1985	DIRECT	No
MASSACHUSETTS	Developing/1994	—	—
MICHIGAN	Developing/1992	DIRECT	No
MINNESOTA	Yes/1991	DIRECT	No
MISSISSIPPI	Yes/1986	DIRECT	SELECT
MISSOURI	Developing/1994	DIRECT	No
MONTANA	Yes/1990	DIRECT	No
NEBRASKA	—	—	—
NEVADA	Yes/1985	No state assessment	RECOMMEND
NEW HAMPSHIRE	—	—	—
NEW JERSEY	Yes/1990	No state assessment	No
NEW MEXICO	Developing/1992	INDIRECT	No
NEW YORK	Yes/1987	DIRECT	No
NORTH CAROLINA	Developing/1995	DIRECT	SELECT
NORTH DAKOTA	Developing/1992	—	No
OHIO	Developing/1993	Developing assessment	No
OKLAHOMA	Revising/1992	DIRECT	No

STATE SCIENCE CURRICULUM FRAMEWORK OR GUIDE			
State	Science Framework or Guide Date of Completion	Framework or Guide Relationship toScience Student Assessment	Framework or Guide Relationship to Science Textbooks
OREGON	Yes 1989	Developing assessment	SELECT
PENNSYLVANIA	Yes 1987	LEARNING OUTCOMES	No
RHODE ISLAND	Developing 1995	No state assessment	No
SOUTH CAROLINA	—	—	—
SOUTH DAKOTA	—	—	—
TENNESSEE	Yes 1990	INDIRECT	SELECT
TEXAS	Yes 1989	LEARNING OUTCOMES	RECOMMEND
UTAH	Developing 1993	DIRECT	SELECT
VERMONT	Developing 1993	No state assessment	—
VIRGINIA	Yes 1986	INDIRECT	SELECT
WASHINGTON	Yes 1991	No state assessment	No response
WEST VIRGINIA	Revising 1992	LEARNING OUTCOMES	SELECT
WISCONSIN	Revising 1986	No state assessment	No
WYOMING	Yes 1990	No state assessement	No
TOTAL	Yes Revising=30 States Developing=15 States No=6 States	DIRECT=16 States INDIRECT=7 States LEARNING OUTCOMES=6 States	SELECT=13 States RECOMMEND=8 States

DIRECT = *Direct linkage between framework or guide and assessment, i. e. curriculum framework defines content topics and skills to be assessed in science.*
INDIRECT = *Curriculum framework defines goal or objectives for instruction, and assessment is developed or selected to reflect goals and objectives.*
LEARNING OUTCOMES = *State has desired learning outcomes, separate from curriculum framework, and the learning outcomes are used to develop the student assessment.*
SELECT = *Science curriculum guide or framework is used to select state-approved textbooks.*
RECOMMEND = *Science curriculum guide or framework is used to recommend a list of textbooks, with selection being made by local districts.*
— *No state curriculum framework or guide.*

Source: State Department of Education, Mathematics and Science Supervisors, Winter, 1992.

Appendix C

STATE TESTS IN SCIENCE AND MATHEMATICS BY GRADE AND TYPE OF TEST

	State Achievement Tests			State Competency(c) & Proficiency (p) Tests		
	Science	Mathematics	Source	Science	Mathematics	Source
ALABAMA	4,8	4,8	Stanford	–	5,6,9(c),11-12(p),Alg I	State
ALASKA	–	4,6,8	ITBS	–	–	–
ARIZONA	–	2-12	ITBS	–	–	–
ARKANSAS	–	4,7,10	Stanford	6,8(c)	5,6,8(c)	State
CALIFORNIA	8	3,6,8,12	State	–	9(p)	State Dist.opt.
COLORADO	–	4,7,10	ITBS TAP('91-92)	–	–	–
CONNECTICUT	4,8,11	4,8,11	State/NAEP	–	4,6,8(c)	State
DELAWARE	–	3,8,11	Stanford	–	8,11(p)	District
D.C.	1-6	3,6,8,9,11	CTBS	–	–	–
FLORIDA	–	4,7,10	4,7 State Dist.opt.	–	11(c)	State
GEORGIA	2,4,7,9	2,4,7,9	ITBS	3,5,8(c),11(p)	3,5,8(c),11(p)	State
HAWAII	–	3,6,8,10	Stanford	–	10-12(c)	State
IDAHO	6,8	6,8	ITBS	–	8(p)	State
ILLINOIS	4,7,11	3,6,8,10	State	–	–	–
INDIANA	–	–	–	3,6,8,11(c)	1,2,3,6,8,9,11(c)	State
IOWA	–	–	–	–	–	–
KANSAS	–	4,7,10	State	–	–	–
KENTUCKY	4,8,12	4,8,12	State	–	K,1,2,3,5,7,10(c)	State
LOUISIANA	4,6,9	3,5,7	CAT	11(p)	5,10(p)	State
MAINE	4,8,11	4,8,11	State	–	–	–
MARYLAND	3,5,8	3,5,8	CTBS	–	9(c)	State
MASSACHUSETTS	4,8,12	4,8,12	State	–	–	–
MICHIGAN	5,8,11	4,7,10	State/NAEP	–	–	–
MINNESOTA	6,9,11	5,8,11	State Dist.opt.	–	–	–
MISSISSIPPI	4,6,8	4,6,8	Stanford	–	5,11(c),Alg I(p)	State
MISSOURI	3,6,8,10	3,6,8,10	State	–	9(c)	State
MONTANA	3,8,11	3,8,11	State Dist.opt.	–	–	–
NEBRASKA	3 levels	3 levels	Dist. opt	–	By 5(c)	Dist.opt.
NEVADA	–	3,6,9	CTBS	–	11-12(p)	State
NEW HAMPSHIRE	–	5('95)	State	–	–	–
NEW JERSEY	–	3,6,8,11	3,6 Dist.opt. State	–	–	–
NEW MEXICO	3,5,8	3,5,8	CTBS	–	10(c)	State
NEW YORK	4	3,6	State	9-12(p),9(c)	9-12(p),9-12(c)	State & Regents
NORTH CAROLINA	3,6,8	3,6,8	State	Courses(p)	3,6,8,10(c),Courses(p)	State
NORTH DAKOTA	3,6,8,11	3,6,8,11	CTBS	–	–	–
OHIO	–	4,6,8,10	State Dist. opt.	–	1-12,(c/p)(p)	Dist.opt./State
OKLAHOMA	3,5,7,9,11	3,5,7,9,11	ITBS TAP	–	–	–
OREGON	–	3,5,8,11	State	–	–	–
PENNSYLVANIA	–	5,8	State	–	–	State
RHODE ISLAND	–	3,6,8,10	MAT	–	–	–

STATE TESTS IN SCIENCE AND MATHEMATICS BY GRADE AND TYPE OF TEST

	State Achievement Tests			State Competancy(c) & Proficiency (p) Tests		
	Science	Mathematics	Source	Science	Mathematics	Source
S.CAROLINA	4,5,7,9,11	4,5,7,9,11	CTBS	3,6,8(c)	1,2,3,6,8,10(c)	State
SOUTH DAKOTA	4,8,11	4,8,11	Stanford	–	–	–
TENESSEE	2-8	2-8	State	–	10(p)	State
TEXAS	–	3-11	State	–	3,5,7,9,11(c)	State
UTAH	–	3-11	CTBS	–	9(p)	State/Dist.opt.
VERMONT	–	–	–	–	–	–
VIRGINIA	4,8,11	4,8,11	ITBS/TAP	–	6-8(c)	State
WASHINGTON	–	4,8,11	CTBS/State	–	–	–
WEST VIRGINIA	–	5,6,9,11	CTBS	–	1-6(c)	State
WISCONSIN	–	–	–	–	3,7,10(c)	State/Dist.opt.
WYOMING	–	–	–	–	–	–
TOTAL	27	46		5(c)+(p)	21(c)15(p)	

Source: State Department of Education, Assessment Directors, Fall, 1991.

APPENDIX D

THE SODA TASK

courtesy of:

CONNECTICUT COMMON CORE OF LEARNING, PERFORMANCE ASSESSMENT PROJECT

SPONSORED BY THE NATIONAL SCIENCE FOUNDATION

DESCRIPTION OF THE TASK

SUMMARY OF THE TASK: Students are asked to identify samples of soda as being diet or regular based on their physical, chemical, and/or biological properties.

DEVELOPED BY: Dale Wolfgram, Jeffrey Greig, Michal Lomask, and Joan Baron.

REVIEWERS: Compton Mahase, Bob Bagioni, Peter Kavali, Jane Knox, George Lelievre, Mike Rollins, Robert Segall, Amy Shively, and CoMPACT III.

COURSE: General Science.

GRADE/LEVEL: 9-12 Medium.

CURRICULUM TOPIC: Physical, chemical, and biological properties. Identification of matter.

PREREQUISITE KNOWLEDGE: Students should have some background knowledge of physical, chemical, and biological properties, and identification of matter. Students should have a background in co-operative group work.

SUGGESTED LENGTH OF TIME: 3-4 class periods.

EQUIPMENT NEEDED: The teacher should display samples of the sodas at the beginning of the task, but should not have any laboratory materials visible until after the students have reached step three in Part II.

Regular Soda	Beakers	Heat Source
Diet Soda	Tripods	Graduated Cylinders
Wire Gauze	Safety Glasses	Aprons

Optional: Yeast, Benedict's Solution, and Glucose Test Strips

References: *Merck Index*, *CRC Handbook of Chemistry and Physics*, chemical dictionary, and chemistry textbooks.

THE SODA TASK

courtesy of

CONNECTICUT COMMON CORE OF LEARNING, PERFORMANCE ASSESSMENT PROJECT

SPONSORED BY THE NATIONAL SCIENCE FOUNDATION

NOTES TO THE TEACHER

SAFETY CONSIDERATIONS

Normal laboratory safety procedures should be followed. Students should wear safety goggles and aprons at all times.

PRIOR PREPARATION

The teacher will have to prepare samples of regular and diet soda and label them A and B. The number of samples should be sufficient for each group to conduct several experiments. The teacher should provide students with varied equipment and materials to support a variety of inquiry methods.

ADMINISTRATION

Students should be given the scoring dimensions in the "Directions to the Students" before beginning any work on the task. Part I is done individually. Students should be given 10-15 minutes to answer the initial question. Part II is done in groups. Students should be given 3-4 class periods to design and carry out their investigations and report their results. Part III is done individually. Students should be given up to 1 class period to complete these final questions.

INFORMATION NEEDED

Some type of clear soda ("-Up, etc.) should be used to obtain the best results. Students should be given an ample supply of soda to complete their tests.

The following describes some of the tests that students may use to distinguish between the two soda samples.

1. Glucose test strips. If students choose to use these test strips, they have to assume that they test for all reducing sugars, not just for glucose.

2. Students may choose to identify the samples based on their density or freezing or boiling points. These tests are valid, although they might not allow for meaningful comparisons due to only small differences in the properties of the two sodas.

3. The use of the "sticky test" will provide students with reliable results since regular soda contains a large amount of sugar while diet soda contains only a small amount of aspartame.

4. The following might serve as possible comparisons:
   a. Adding salt to the diet soda causes more fizziness than the regular soda.
   b. Conductivity of the two sodas.
   c. Diet soda may have a stronger aroma than the regular soda.
   d. Students may observe that the two sodas differ in color or amount of fizz. Diet soda may have more and larger bubbles.

NOTES TO THE TEACHER (Continued)

5. Adding yeast to the two sodas might show a difference. The yeast will metabolize the sugars in the regular soda to produce the energy they need. During this process, the yeast will break down the sugars into carbon dioxide, which is released from the water in visible gas bubbles. Gentle warming will accentuate this. This test will work only with yeast that can't metabolize aspartame as a source of energy.

6. The benedict's solution test will be positive for some regular sodas, producing a reddish precipitate with slight warming. Diet soda containing fructose will give the same result. Sucrose is not a reducing sugar and therefore will not react with the benedict solution. Aspartame is also not a reducing agent and therefore will not react with the benedict solution.

7. Sulfuric acid will react with reducing sugars to produce a caramel.

8. More sugar will dissolve in diet soda than in regular soda.

9. If the two sodas are partially evaporated, the regular soda will leave more residue. (If students evaporate the soda completely, a black residue will be left that may be difficult to remove from glassware.)

10. All comparisons of unknown samples to the characteristics of known samples should be considered as one testing method. Students should be asked to perform another test as well.

GUIDANCE

No guidance should be given to students in the design and implementation of their investigation, other than to check that students are following proper safety procedures. Students should always show their proposed plans to the teacher before carrying out their experiments due to the open-endedness of the task.

SCORING

The scoring objectives and criteria can be found on the Objectives Rating Forms for Group and Individual.

THE SODA TASK

courtesy of

CONNECTICUT COMMON CORE OF LEARNING, PERFORMANCE ASSESSMENT PROJECT

SPONSORED BY THE NATIONAL SCIENCE FOUNDATION

INSTRUCTIONS TO THE STUDENT

Part I: Getting Started by Yourself

You will be given two samples of soda, one regular soda containing sugar and the other one diet soda containing an artificial sweetener. Your task is to identify each sample as diet or regular based on your knowledge of physics, chemistry, and/or biology. As in any experiment, you are not allowed to taste any of the samples.

Information about Regular and Diet Soda

Regular soda generally contains fructose and/or glucose (types of sugar) as sweeteners. Diet soda generally contains aspartame as a sweetener. The chemical formulas for these ingredients are shown below:

Fructose $C_6H_{12}O_6$
Sucrose $C_{12}H_{22}O_{11}$
Aspartame $C_{14}H_{18}N_2O_5$

Aspartame is roughly one hundred times sweeter than sugar; therefore, significantly less aspartame is needed to make a given amount of diet soda equally sweet as regular soda.

Make a list of the properties of the two sodas which might help to distinguish between the samples. Write down as many as you can think of.

Part II: Group Work

1. Make a group list of the properties of the two sodas which might help to distinguish between the samples.
2. Based on your list of properties, design two tests to distinguish between the two types of soda. They should be the ones which your group believes would be the most effective in distinguishing between the two samples. Explain why you chose each of them. Show that you understand the science involved in each test.
3. Write out a complete experimental plan for each of the two tests. Include a list of all the materials and equipment that you will need. Show your plan to your teacher before proceeding.

After getting approval from your teacher, carry out your experiments.

4. Summarize your group's findings in a final report which includes:
   a. What your group tried to investigate (dependent and independent variables).
   b. How your group performed your experiments (method).
   c. What your group found (raw data, organized in charts or graphs, as necessary).
   d. What your group concluded (based on experimental findings) and how valid your group thinks these conclusions are (including sources of error).
5. Prepare an oral presentation of your group's experiments, findings, and conclusions. Each member of your group should be ready to participate in any part of the presentation.
6. After hearing all of the oral presentations, answer the following question: If you were a diabetic and had to know whether a sample of soda had sugar in it, which test would your group trust the most? Which test would your group trust the least? Explain fully why you chose each of these.

THE SODA TASK

courtesy of

CONNECTICUT COMMON CORE OF LEARNING, PERFORMANCE ASSESSMENT PROJECT

SPONSORED BY THE NATIONAL SCIENCE FOUNDATION

INSTRUCTIONS TO THE STUDENT (continued)

Part III: Finishing by Yourself

1. If you were given two samples of water, one of which is salt water and the other fresh water, which tests that your class tried out for the sodas would be useful in differentiating between the two? Which tests would not be useful? What other new tests that your class did not try might be appropriate for this problem?
2. The following report was completed by one group of students working on "The Soda Task." Read the report and answer the questions that follow.

Group Report

Our group tested the following two properties of the sodas:

Test #1: Boiling

The boiling point of soda A was 96 and soda B was 97. The higher sugar content of B must have increased its boiling point.

Test #2: Density

Procedure: Weigh graduated cylinder. Measure 100 mL of soda A and weigh the cylinder and soda together.

Data: Mass of graduated cylinder = 43.26 g
Mass of cylinder and soda A = 141.45 g
Mass of cylinder and soda B = 144.02 g

Analysis: Density = mass/volume
Density of soda A = (141.45 - 43.26)g/100 mL = .9819 g/mL
Density of soda B = (144.01 - 43.26)g/100 mL = 1.0075 g/mL
Soda A was less dense than soda B.

Final Conclusion: Due to the observations from the boiling test and the calculated density, Soda A was diet soda and soda B was regular soda.

3. a. A scientific report is written to share information and to enable others to replicate (repeat) the same experiment. Does this report give you enough information to replicate the experiment? If not, what is missing or not completely described in the report? Please be specific in your critique.

   b. Do you think this group's conclusion is valid? Explain fully why you think so.

SCORING GUIDE

The Soda Task - Dimension II: Group Experimentation

Items below refer to Instructions to the Student, Part II: Group Work	**Excellent**	**Good**	**Fair**	**Poor**
II. 1 Identification of properties.	❒ 3 or more	❒ 2 properties	❒ 1 property	❒ 0 properties
1. Chemical ingredients 3. Boiling freezing point 5. Conductivity 2. Density 4. Solubility Others:				
II. 2 Experimental design.	❒ All 4	❒ 3	❒ 2	❒ 0-1
The experimental design should: 1. match the factor to be studied. 2. define independent and dependent variables. 3. control and test variables separately. 4. be clearly described.				
II. 3 Performance of experiments.	Test 1: ❒ Yes ❒ No		Test 2: ❒ Yes ❒ No	
Yes - Indication that students have either attempted to control variables or have considered how this might affect their results. No - No indication of the above.				
II. 4 Data collection and organization.	❒ All 6	❒ 4-5	❒ 2-3	❒ 0-1
Quality of measurements: 1. Accuracy of data. 2. Repetition of experiments (until data are replicated). *Use of mathematics:* 3. Clarity and organization (e.g., proper labels, units, scaling, etc.). 4. Appropriate symbolic representation (e.g., use of bar graphs vs. Cartesian coordinate graphs). *Manipulation and presentation of data:* 5. Making calculations (e.g., taking averages). 6. Correct use of formulas to define new terms (e.g., density)—when appropriate.				

BEST COPY AVAILABLE

The Soda Task - Dimension II: Group Experimentation (continued)	Excellent Good Fair Poor
II. 5 Conclusions.	Test 1: ❒ Yes ❒ No Test 2: ❒ Yes ❒ No
Yes - Conclusions made are consistent with and supported by the data collected and are valid (based on accurate scientific explanation). No - Conclusions made are not consistent with or supported by the data collected or are invalid. If conclusions are discrepant, students should resolve the conflict.	
II. Reflection: Identification of most and least trusted tests.	❒ #1 and #2 ❒ #1 or #2 ❒ #3 ❒ #4
1. Trustworthy test selected with full explanation. 2. Non-trustworthy test selected with full explanation. 3. Trustworthy and non-trustworthy tests selected with little or no explanation. 4. Inappropriate tests selected with little or no explanation.	

BEST COPY AVAILABLE

SCORING GUIDE (continued)

The Soda Task - Dimension I: Individual Understanding

Items below refer to Instructions to the Student, Part III: Finishing by Yourself	Excellent	Good	Fair	Poor
III. 1 Select most and least trusted tests based on their scientific validity.	☐ All 3	☐ 2	☐ 1	☐ 0
The following should be included: 1. Useful method with explanation. 2. Nonuseful method with explanation. 3. New useful method with explanation.				
III. 2 Analysis of research information.	☐ All 4	☐ 2	☐ 1	☐ 0
The following deficiencies of the group report should be identified: 1. No description of method given for determining boiling point test. 2. No units of temperature are included for the boiling point measurements. 3. Convention for significant figures not allowed.				
III. 3 Analysis of research conclusions.	☐ All 3	☐ 2	☐ 1	☐ 0
Conclusions should be seen as questionable due to the following: 1. No explanation given as to how the conclusions were made (why higher sugar content results in higher boiling point or why the regular soda should be more dense). 2. Uncertainty about the accuracy of the measurements. 3. No repetition of trials is indicated.				

BEST COPY AVAILABLE

A Brief Guide to ERIC

The Educational Resources Information Center
Office of Educational Research and Improvement
U.S. Department of Education

What is ERIC?

The Educational Resources Information Center (ERIC) is a national education information network designed to provide users with ready access to an extensive body of education-related literature. Established in 1966, ERIC is supported by the U.S. Department of Education, Office of Educational Research and Improvement.

The ERIC database, the world's largest source of education information, contains over 735,000 abstracts of documents and journal articles on education research and practice. This information is available at more than 2,800 libraries and other locations worldwide.

You can access the ERIC database by using the print indexes *Resources in Education* and *Current Index to Journals in Education*, online search services, or CD-ROM at many libraries and information centers. The database is updated monthly (quarterly on CD-ROM).

The ERIC System

The ERIC System, through its 16 subject-specific Clearinghouses, 4 Adjunct Clearinghouses, and four support components, provides a variety of services and products that can help you stay up to date on a broad range of education-related issues. Products include research summaries, publications on topics of high interest, newsletters, and bibliographies. ERIC system services include computer search services, reference and referral services, and document reproduction. ACCESS ERIC, with its toll-free number, 1-800-LET-ERIC, informs callers of the services and products offered by ERIC components and other education information service providers.

ERIC Reference and Referral Services

With the world's largest educational database as a resource, ERIC staff can help you find answers to education-related questions, refer you to appropriate information sources, and provide relevant publications. ERIC components answer more than 100,000 inquiries each year. Questions should be directed to ACCESS ERIC or a specific Clearinghouse.

Specific documents: Requests for documents in the ERIC database for which you have an accession number (ED number) should be referred to an information provider near you. Call ACCESS ERIC to locate the nearest ERIC education information provider.

Subject-specific topics: Subject-related questions should be directed to the particular ERIC Clearinghouse whose scope is most closely associated with the subject matter involved. Or, call ACCESS ERIC for a referral.

Computer searches: Requests for a computer search should be directed to one of the search services listed in the *Directory of ERIC Information Service Providers*, available from ACCESS ERIC.

ERIC Clearinghouse publications: Requests for a publication produced by an ERIC Clearinghouse should be directed to the specific Clearinghouse.

Major ERIC Products

ERIC produces many products to help you access and use the information in the ERIC database:

Abstract Journals: ERIC produces two monthly abstract journals. *Resources in Education* (RIE), a publication announcing recent education-related documents, and the *Current Index to Journals in Education* (CIJE), a periodical announcing education-related journal articles, is available through Oryx Press (1-800-457-6799). Many libraries and information centers subscribe to both monthly journals.

***All About ERIC*:** This guide provides detailed information on ERIC, its products and services, and how to use them. Free copies are available from ACCESS ERIC.

***Catalog of ERIC Clearinghouse Publications*:** The *Catalog* lists publications produced by the ERIC Clearinghouses and support components, prices, and ordering information. Free copies of the *Catalog* are available from ACCESS ERIC.

***The ERIC Review*:** This journal discusses important ERIC and education-related developments. For a copy, call ACCESS ERIC.

Information Analysis Products: ERIC Clearinghouses produce reports, interpretive summaries, syntheses, digests, and other publications, many free or for a minimal fee. Contact the Clearinghouse most closely associated with your interests for its publications list. Call ACCESS ERIC for a free copy of the *Catalog of ERIC Clearinghouse Publications*.

Microfiche: The full text of most ERIC documents is available on microfiche. Individual documents and back collections on microfiche are available. Call the ERIC Reproduction Document Service (EDRS) for more information.

Thesaurus of ERIC Descriptors - The complete list of index terms used by the ERIC System, with a complete cross-reference structure and rotated and hierarchical displays, is available from Oryx Press.

ERICTAPES - Computer tapes of the ERIC database are available by subscription or on demand from the ERIC Facility (write for a price list).

ERIC Document Delivery

Documents: EDRS is the primary source for obtaining microfiche or paper copies of materials from the ERIC database. EDRS can provide full-text copies of most documents announced in Resources in Education (RIE), and ERIC's microfiche collection is available by monthly subscription from EDRS. EDRS also sells microfiche and paper copies of individual documents on request. For more information, call EDRS at (800) 443-ERIC.

Journal Articles: Two agencies that provide reprint services of most journal articles announced in *Current Index to Journals in Education* (CIJE) are listed below. Some journals do not permit reprints; consult your local university or local library to locate a journal issue. Or, write directly to the publisher. Addresses are listed in the front of each CIJE.

University Microfilms International (UMI)
Article Clearinghouse
300 North Zeeb Road
Ann Arbor, MI 48106
Telephone: (800) 732-0616

Institute for Scientific Information (ISI)
Genuine Article Service
3501 Market Street
Philadelphia, PA 19104
Telephone: (800) 523-1850

ERIC Information Retrieval Services

The ERIC database is one of the most widely used bibliographic databases in the world. Last year, users from 90 different countries performed nearly half a million searches of the database. The ERIC database currently can be searched via four major online and CD-ROM vendors (listed below). Anyone wishing to search ERIC online needs a computer or terminal that can link by telephone to the vendor's computer, communications software, and an account with one or more vendors.

The *Directory of ERIC Information Providers* lists the address, telephone number, and ERIC collection status for more than 900 agencies that perform searches. To order a copy, call ACCESS ERIC (1-800-LET-ERIC).

Online Vendors

BRS Information Technologies
8000 Westpark Drive
McLean, VA 22102
Telephone: (703) 442-0900
(800) 289-4277

Dialog Information Services
3460 Hillview Avenue
Palo Alto, CA 94304
Telephone: (415) 858-2700
(800) 334-2564

OCLC (Online Computer Library Center, Inc.)
6565 Frantz Road
Dublin, OH 43017-0702
Telephone (614) 764-6000
(800) 848-5878 (Ext. 6287)

CD-ROM Vendors

Dialog Information Services (same address as above)

Silver Platter Information Services
One Newton Executive Park
Newton Lower Falls, MA 02162-1449
Telephone: (617) 969-2332
(800) 343-0064

ERIC Components

Federal Sponsor

Educational Resources Information Center (ERIC)
U.S. Department of Education
Office of Educational Research and Improvement (OERI)
555 New Jersey Avenue N.W.
Washington, DC 20208-5720
Telephone: (202) 219-2289
Fax: (202) 219-1817

Clearinghouses

Dr. Susan Imel, Director
ERIC Clearinghouse on Adult, Career, & Vocational Education
CETE/The Ohio State University
1900 Kenny Road
Columbus, OH 43210-1090
Telephone: (614) 292-4353; (800) 848-4815
Fax (614) 292-1260
Internet: ericacve@magnus.acs.ohio-state.edu

Dr. Lawrence M. Rudner, Director
ERIC Clearinghouse on Assessment and Evaluation
The Catholic University of America
Department of Education
209 O'Boyle Hall
Washington, DC 20064
(202) 319-5120
Internet: eric_ae@cua.edu

Dr. Arthur M. Cohen, Director
ERIC Clearinghouse for Community Colleges
University of California at Los Angeles (UCLA)
3051 Moore Hall
Los Angeles, CA 90024-1521
Telephone: (310) 825-3931; (800) 832-8256
Fax: (213) 206-8095
Internet: eeh3usc@mvs.oac.ucla.edu

Dr. Garry R. Walz, Director
ERIC Clearinghouse on Counseling and Student Services
University of North Carolina at Greensboro
School of Education
1000 Spring Garden Street
Greensboro, NC 27412-5001
Telephone: (919) 334-4114
Fax: (919) 334-4116
Internet: bleuerj@iris.uncg.edu

Dr. Bruce A. Ramirez, Director
ERIC Clearinghouse on Disabilities and Gifted Education
Council for Exceptional Children
1920 Association Drive
Reston, VA 22091-1589
Telephone: (703) 264-9474; (800) 328-0272
Fax: (703) 264-9494
Internet: ericec@inet.ed.gov

Dr. Philip K. Piele, Director
ERIC Clearinghouse on Educational Management
University of Oregon
1787 Agate Street
Eugene, OR 97403-5207
Telephone: (503) 346-5043; (800) 438-8841
Fax: (503) 346-5890
Internet: ppiele@oregon.uoregon.edu

Dr. Lilian Katz, Director
ERIC Clearinghouse on Elementary & Early Childhood Education
University of Illinois
805 W. Pennsylvania Avenue
Urbana, IL 61801-4897
Telephone: (217) 333-1386; (800) 583-4135
Fax (217) 333-5847
Internet: ericeece@ux1.cso.uiuc.edu

Dr. Jonathon D. Fife, Director
ERIC Clearinghouse on Higher Education
The George Washington University
One Dupont Circle N.W., Suite 630
Washington, DC 20036-1183
Telephone: (202) 296-2597
Fax: (202) 296-8379
Internet: eriche@inet.ed.gov

Dr. Michael B. Eisenberg, Director
ERIC Clearinghouse on Information & Technology
Syracuse University
4-194 Center for Science and Technology
Syracuse, NY 13244-4100
Telephone: (315) 443-3640; (800) 464-9107
Fax: (315) 443-5732
Internet: eric@ericir.syr.edu
AskERIC (Internet-based question-answering service):
askeric@ericir.syr.edu

Dr. Charles W. Stansfield, Director
ERIC Clearinghouse on Languages and Linguistics
Center for Applied Linguistics (CAL)
1118 22nd Street N.W.
Washington, DC 20037-0037
Telephone: (202) 429-9551 and (202) 429-9292
Fax: (202) 429-9766 and (202) 659-5641
Internet: cal@guvax.georgetown.edu
**Includes Adjunct ERIC Clearinghouse on Literacy Education for Limited English Proficient Adults*

Dr. Carl B. Smith, Director
ERIC Clearinghouse on Reading, English, and Communication
Indiana University
Smith Research Center (SRC), Suite 150
2805 East 10th Street
Bloomington, IN 47408-2698
Telephone: (812) 855-5847; (800) 759-4723
Fax: (812) 855-7901
Internet: ericcs@ucs.indiana.edu

Mr. Craig Howley, Director
ERIC Clearinghouse on Rural Education and Small Schools
Appalachia Educational Laboratory (AEL)
1031 Quarrier Street
P.O. Box 1348
Charleston, WV 25325-1348
Telephone: (304) 347-0400; (800) 624-9120
Fax: (304) 347-0487
Internet: u56d9@wvnvm.wvnet.edu

Dr. David Haury, Director
ERIC Clearinghouse on Science, Mathematics, and Environmental Education
The Ohio State University
1929 Kenny Road
Columbus, OH 43210-1080
Telephone: (614) 292-6717
Fax (614) 292-0263
Internet: ericse@osu.edu

Dr. John Patrick, Director
ERIC Clearinghouse on Social Studies/Social Science Education**
Indiana University
Social Studies Development Center (SSDC)
2805 East 10th Street, Suite 120
Bloomington, IN 47408-2698
Telephone: (812) 855-3838; (800) 266-3815
Fax: (812) 855-7901
Internet: ericso@ucs.indiana.edu
***Includes Adjunct ERIC Clearinghouse on Art Education; and the National Clearinghouse for U. S.-Japan Studies*

Dr. Mary Dilworth, Director
ERIC Clearinghouse on Teaching and Teacher Education
American Association of Colleges for Teacher Education (AACTE)
One Dupont Circle N.W., Suite 610
Washington, DC 20036-1186
Telephone: (202) 293-2450
Fax: (202) 457-8095
Internet: jbeck@inet.ed.gov

Dr. Erwin Flaxman, Director
ERIC Clearinghouse on Urban Education
Teachers College, Columbia University
Institute for Urban and Minority Education
Main Hall, Room 303, Box 40
525 West 120th Street
New York, NY 10027-9998
Telephone: (212) 678-3433; (800) 601-4868
Fax (212) 678-4048
Internet: eric-cue@columbia.edu

Adjunct Clearinghouses

Chapter 1
Chapter 1 Technical Assistance Center
2601 Fortune Circle East
One Park, Fletcher Building, Suite 300-A
Indianapolis, IN 46241-2237
Toll Free: (800) 456-2380
Telephone: (317) 244-8160
Fax: (317) 244-7386

Clinical Schools
American Association of Colleges for Teacher Education
One Dupont Circle, NW, Suite 610
Washington, DC 20036-1186
Telephone: (202) 293-2450
Internet: iabdalha@inet.ed.gov

Consumer Education
National Institute for Consumer Education
207 Rackham Building, West Circle Drive
Eastern Michigan University
Ypsilanti, MI 48197-2237
Toll Free: (800) 336-6423
Telephone: (313) 487-2292
Internet: cse_bonner@emunix.emich.edu

ESL Literacy Education
Center for Applied Linguistics
1118 22nd Street NW
Washington, DC 20037
Telephone: (202) 429-9292 (ext. 200)
Internet: cal@guvax.georgetown.edu

Law-Related Education
Indiana University
Social Studies Development Center
2805 East 10th Street, Suite 120
Bloomington, IN 47408-2698
Toll Free: (800) 266-3815
Telephone: (812) 855-3838
Internet: ericso@ucs.indiana.edu

Test Collection
Rosedale Road
Princeton, NJ 08541
Telephone: (202) 319-5120
Internet: eric@ae@cua.edu

U. S.-Japan Studies
Indiana University
Social Studies Development Center
2805 East 10th Street, Suite 120
Bloomington, IN 47408-2698
Fax: (812) 855-7901

Support Components

ACCESS ERIC
Aspen Systems Corporation
1600 Research Boulevard
Rockville, MD 20850-3166
Telephone: (800) LET-ERIC
Fax: (301) 251-5212

ERIC Document Reproduction Service
7420 Fullerton Road, Suite 110
Springfield, VA 22153-2852
Telephone: (301) 258-5500; (800) 443-ERIC
Fax: (301) 948-3695

ERIC Processing and Reference Facility
1301 Piccard Drive
Rockville, MD 20850-4305
Telephone: (301) 258-5500
Fax: (301) 948-3695

Oryx Press
4041 North Central Ave., Suite 700
Phoenix, AZ 85012-3399
Telephone: (602) 265-2651; (800) 279-ORYX
Fax: (602) 265-6250; (800) 279-4663

How to Submit Documents to ERIC

ERIC collects a variety of materials on education-related topics. Examples of materials included in the database:

Research reports
Instructional materials
Monographs
Teaching Guides
Speeches and presentations
Manuals and handbooks
Opinion papers

Submissions can be sent to the Acquisitions Department of the ERIC Clearinghouse most closely related to the subject of the paper submitted, or sent to the ERIC Processing Facility.

About the Author...

Margaret Jorgensen received her Ph.D in measurement, evaluation, and statistical analysis from the University of Chicago and has worked in the field of assessment in both theoretical and applied areas for twenty years. She is currently a Senior Examiner in the Southern Field Office of Educational Testing Service.

BEST COPY AVAILABLE